Anne Reboul *Editor*

Mind, Values, and Metaphysics

Philosophical Essays in Honor of Kevin Mulligan - Volume 2

Springer

Mind, Values, and Metaphysics

Anne Reboul
Editor

Mind, Values, and Metaphysics

Philosophical Essays in Honor of Kevin Mulligan - Volume 2

Springer

Editor
Anne Reboul
Institute for Cognitive Sciences
Bron Cedex
France

ISBN 978-3-319-05145-1 ISBN 978-3-319-05146-8 (eBook)
DOI 10.1007/978-3-319-05146-8
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014941197

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Contributors

Bernard Baertschi Department of philosophy, University of Geneva, Geneva, Switzerland

Jonathan Barnes Université Paris-Sorbonne, Paris, France

Clotilde Calabi Universita degli Studi di Milano, Milano, Italy

Roberto Casati Institut Nicod, CNRS (EHESS-ENS), Paris, France

Laurent Cesalli CNRS, Université de Lille 3 (UMR 8163: Savoirs, Textes, Langage), Université de Genève, Geneva, Switzerland

Julien A. Deonna University of Geneva, Geneva, Switzerland

Jérôme Dokic Institut Jean-Nicod (CNRS, EHESS, ENS), Paris, France

Pascal Engel Université de Genève, Genève, Suisse

Manuel García-Carpintero LOGOS-Departament de Lògica, Història i Filosofia de la Ciència, Universitat de Barcelona, Barcelona, Spain

Philip Gerrans University of Adelaide, Adelaide, Australia

Anita Konzelmann Ziv University of Geneva, Geneva, Switzerland

Wolfgang Künne University of Hamburg, Hamburg, Germany

Patrizia Lombardo Department of French and Center for Affective Sciences CISA, University of Geneva, Geneva, Switzerland

Diego Marconi University of Turin, Turin, Italy

Uwe Meixner University of Augsburg, Augsburg, Germany

Carlos J. Moya University of Valencia, Valencia, Spain

Bence Nanay University of Cambridge, Cambridge, UK

University of Antwerp, Antwerp, Belgium

Martine Nida-Rümelin Université de Fribourg, Fribourg, Switzerland

Alfredo Paternoster University of Bergamo, Bergamo, Italy

Anne Reboul Laboratory on Language, the Brain and Cognition (L2C2), CNRS UMR5304, Institute for Cognitive Sciences-Marc Jeannerod, Lyon, France

Anne Reboul Laboratoire sur le Langage, le Cerveau et la Cognition (L2C2 CNRS UMR5304), Institut des Sciences Cognitives Marc Jeannerod, Bron cedex, France

François Recanati Institut Jean-Nicod, CNRS-ENS-EHESS, Paris, France

Luigi Rizzi University of Siena, Sienna, Italy

University of Geneva, Geneva, Switzerland

Toni Rønnow-Rasmussen Lund University, Lund, Sweden

Marco Santambrogio Università di Parma, Parma, Italy

Göran Sundholm Leiden University, Leiden, Netherland

Christine Tappolet Université de Montréal, Montréal, Canada

Fabrice Teroni University of Bern, Bern, Switzerland

Cain Todd Lancaster University, Lancaster, UK

University of Fribourg, Fribourg, Switzerland

Maria van der Schaar Institute for Philosophy, Leiden University, Leiden, Netherlands

Alberto Voltolini University of Turin, Turin, Italy

Chapter 1
Introduction

Anne Reboul

The present volume originates in a collection of papers presented to Kevin Mulligan in 2011 to celebrate his 25 years of professorship in Geneva. It is the second volume of a two-volume set reproducing those papers. The contributors have written papers addressing the main topics Mulligan himself has been interested in during his career.

The first volume is dedicated to two main fields: *metaphysics,* with a specific interest in truth-makers, tropes and relations, and *history of philosophy,* with an emphasis on Austrian philosophy. The second volume gathers chapters on *ethics, values and emotions,* on *epistemology*, *perception and consciousness,* as well as *philosophy of mind and philosophy of language*.

In this introduction, I will briefly indicate the gist of each chapter in the present second volume.

1.1 Part I: Values, Ethics and Emotions

Carlos Moya, in 'Alternatives and Responsibility: an Asymmetrical Approach', discusses the principle of alternative possibilities (PAP) and the intuition that alternative possibilities are necessary for moral responsibility. He defends an asymmetrical view, according to which PAP is true for blameworthy but not for praiseworthy moral responsibility. Moya bases his asymmetric view on the notion of moral obligation, grounded in social institutions. It is linked to the fact that we ask whether the agent *could* have done otherwise only when she *should not* have acted as she did. Moya introduces a distinction between 'Luther' cases—where the agent not only acted as she should, but knows that she could not have acted otherwise—and 'Frankfurt' cases, where the agent is responsible for her action, though, unknown to her, she could not have done otherwise. Both Luther and Frankfurt cases have been used against the PAP and its relation to moral responsibility. Moya first examines Luther cases, which show that the PAP is not true for praiseworthy actions.

A. Reboul (✉)
Laboratory on Language, the Brain and Cognition (L2C2), CNRS UMR5304,
Institute for Cognitive Sciences-Marc Jeannerod, Lyon, France
e-mail: reboul@isc.cnrs.fr

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_1,

However, when an agent performs a blameworthy action, we will not rest content with her belief that this was the only way she could have acted. Here, alternative possibilities come in as 'unfulfilled moral obligations that the agent was able to fulfil'. In other words, we assume that when an agent *should* have acted otherwise, she *could* have done so (an instance of 'ought' implies 'can' or OIC). Hence the asymmetry. Moya defends OIC against its critics. PAP can be derived from OIC, but only in the case of blameworthiness. He then turns to Frankfurt cases. According to him, they should be assessed from the standpoint of moral responsibility, that is, from the view that the agent should have done otherwise. This again justifies the asymmetry whereby the PAP only applies to blameworthy actions, given that the same reasoning does not apply to praiseworthy actions. Adversaries of the PAP, basing their arguments on Frankfurt cases, tend to consider the use of OIC as a case of begging the question. Therefore, Moya proposes to replace the OIC by another principle, the 'Doing Everything one Can' (DEC), that is more intuitively correct and that enjoins the agent to do everything she can reasonably do to fulfil her moral obligations, but does not hold her as morally blameworthy for not doing more. He shows that this is enough to dispose of putative counterexamples based on Frankfurt cases and concludes that the asymmetry stands.

In Chap. 3, Christine Tappolet discusses 'The Normativity of Evaluative Concepts'. Tappolet notes that, while normative concepts seem fairly heterogeneous, one can separate them into two groups: evaluative or axiological (*good–bad*) on one side, deontic (*right–wrong*) on the other. This distinction raises the question of the relation between these two classes of normative concepts, notably that of conceptual priority. Tappolet is mainly interested in the question of whether evaluative concepts are normative, a question which is all the more pressing given the division between evaluative and deontic concepts. Both concepts in the evaluative and concepts in the deontic groups are linked by direct *internal* inferential links, while inferential relations *between* the two groups seem slacker. Additionally, the evaluative class of concepts is much bigger than is the deontic one, the evaluative, but not the deontic, concepts are associated with affective reactions, and evaluative concepts can take comparative and superlative forms, while deontic concepts cannot. All deontic judgments can be transformed into evaluative judgments while the reverse is not true of all evaluative judgments; likewise, evaluative terms have a much wider range of applications than have deontic concepts. Finally, deontic judgments give rise to authentic dilemmas, while evaluative judgments do not. Thus, there is no doubt that the two classes are distinct and, given the obvious link between deontic concepts and normativity, the question arises whether we have any reason to see evaluative concepts as normative. A first way of answering this question is through the view that *ought* is the central normative concept. This leads to the suggestion that what is *good* is also what *ought to be*. However, though it works for the thin evaluative concepts, such as *good,* it is not clear that it would work for the thick evaluative concepts, such as *courageous*. Additionally, though what one *ought to do* is normative, it is not clear that what *ought to be* is. The relevant principle (understood in a non-consequentialist way) here is that one ought to do both what one can do and what is the best (what ought to be). Another link between evaluative con-

cepts and *ought* lies in the appropriateness of our affective reactions towards values. An alternative approach lies in linking normativity and reason, a move which can do justice to thick as well as thin evaluative concepts. It should be noted, nevertheless, that the notion of reason is linked to the concept *ought*. Tappolet concludes that the web of relations between evaluative and deontic concepts suggests that they both belong to the same conceptual level, where the normative applies.

In Chap. 4, 'For Kevin's Sake', Toni Rønnow-Rasmussen discusses the idiom 'for someone's sake' relative to the distinction between impersonal versus personal values. The distinction goes between what is good *simpliciter* and what is good for someone or for someone's sake. Rønnow-Rasmussen notes that the 'fitting-attitude' analysis of values gives pride of place to personal values. He begins by distinguishing three attitudes towards the distinction: Moorean monists insist that there are only impersonal values and that personal values can be reduced to impersonal values; Hobbesian monists reject the very existence of impersonal values and hold that there are only personal values; finally, value dualists admit the existence of both personal and impersonal values. Rønnow-Rasmussen embraces a dualist account and begins by delineating the difference upon which the distinction is based. He suggests that impersonal values are the simple case, where an entity is valued for its own sake, while, in personal values, an entity is also, additionally, valued for someone's sake. This view, however, meets with the non-translatability objection, which rests on the fact that the idiom *for X's sake* is not found in all languages. Rønnow-Rasmussen gives a few translations in a number of languages, showing that 'sake' is ambiguous between an evaluative and a nonevaluative interpretation. Rønnow-Rasmussen then turns to another potential objection to the fitting-attitude analysis, the problem raised when two entities are equally valuable for their own sakes but heterogeneous in their values for the sake of different individuals. On a dualist analysis, this is not paradoxical, however.

Anita Konzelmann Ziv, in 'Knowledge, Emotion, Value and Inner Normativity: KEVIN Probes Collective Persons', is concerned with how personal, affective-based normativity can be attributed to 'social' or 'plural' persons. Social ontology acknowledges that some social groups have personality or are person-like, that they have essential features of persons (such as intentionality and rationality) and are discontinuous with the individual mental states of their constituent members. On the other hand, inner or personal normativity is deeply axiological, linking personality and axiological commitments and behaviours. Konzelmann Ziv uses the acronym KEVIN (knowledge, emotion, value, inner normativity) for this axiological conception of the person. She notes that ascribing a collective ethos to a group may not be enough to see it as an axiological person, unless groups are considered as continuous to their individual members on an affective level, which sounds unlikely. Beyond the question whether KEVIN could accommodate collective persons, there is the question whether it has to do so. KEVIN views an ethos as manifested in a person's life in both a nonnormative way (through her tendencies) and a normative way, through a *vocation* that exerts 'valuational pulls' on the person, without being the product of her conscious volition. Thus, KEVIN intimately links ethos, vocation and person, making it impossible to attribute ethos to anything but a person.

Substituting the notion of conscience to that of vocation does not solve the problem, because conscience seems more limited (to morals) than vocation, and because it requires a noncontigent continuity between those person-relevant properties that are collective and those that are individual. The primary valuation in personalism is preferences (not choices) and axiological knowledge is not propositional. In other words, rather than explain actions through desires and beliefs, KEVIN explains them through axiological knowledge and emotional attitudes. The defender of collective persons can involve speech acts (which are within the abilities of such entities), allowing for self-ascription of aggregate attitudes. This, however, does not open the door to axiological concerns, and Konzelmann Ziv turns to Scheler's view on collective valuations in terms of a consensual (rather than compromise) aggregate of individual attitudes, reached through the twin-person thesis that persons are both 'intimate' and 'social'. Ethos cuts across the intimate/social distinction inside the notion of a person. The consensus theory proposed by Scheler is a promising solution to the necessary integration of individual values into the collective ethos that characterizes collective persons.

Bernard Baertschi deals with 'The Argument of Ethical Naturalism' in Chap. 6. Naturalism in ethics seems caught in a dilemma between charges of naturalistic fallacy, on the one hand, and the indubitable fact that ethics, since Hume, has routinely made reference to human nature. Baertschi aims at defending ethical naturalism in a version that claims that normative justification can legitimately invoke human nature. The charge of naturalistic fallacy is that it is not possible to derive norms and values from facts about human nature. As Baertschi notes, the relevant notion of fact is that found in the 'fact-value' dichotomy, i.e. the 'is-ought' question. Here, the relevant question is whether there are natural ends we ought to pursue. If natural ends are defined in the sense of the theory of natural selection, a positive answer seems unlikely and, compared with other candidates for justification of ethical norms, such as will and reason, human nature seems unpromising. However, this depends on how 'human nature' is understood and moral sentiments such as sympathy and benevolence, as well as other emotions, are rather better candidates. Indeed, ethical naturalism can be based on an essentialist-teleological Aristotelian view, according to which one should distinguish between the factual and the essential natures of man and this essential nature is our *telos* or end. While this might not be enough to accommodate moral duties towards others, an appeal to moral sentiments can complete the picture. Baertschi's conclusion is that ethical naturalism is in no worse position than other ethical theories when it comes to justify norms of conduct.

In Chap. 7, Cain Todd explains 'Why We Don't Perceive Aesthetic Properties'. While aesthetic properties seem to be perceived, our access to them is distinguished from perception by its opacity and by its nonattributive phenomenology. The view that the access to aesthetic properties is perceptual in nature has been common in philosophy, from Hume on, though it is usually taken to involve training. This strongly implies that aesthetic perceptual judgments involve *cognitive penetration,* that is, the top-down influence of beliefs, desires and other mental states on perception. Todd notes that strong intentionalist or representationalist theories of percep-

tion insist on the *transparency* of perception. The reason why aesthetic experiences are not entirely transparent is that they include, in addition to the representational elements, evaluative elements that are not representational. Additionally, given that the representationalist character of ordinary perception means that it is *attributive* (the redness is attributed to the tomato, not to our perception of it), the evaluative elements in aesthetic judgments are not, given that they are subject relative and indeed dependent on imagination. In other words, aesthetic judgments can be seen as 'construals'. This leads Todd to a discussion of the phenomenology of attention relative to the phenomenology of perception: The former lacks the 'phenomenology of objectivity' that the second has, the salience resulting from attention being a property of the experience rather than of its object. Todd argues that this nonattributivity and opacity of attention can be found in aesthetic perceptual judgments, which is no accident as there is a strong chance that attention is indeed involved in aesthetic perceptual judgment. The more we are aware of our own contribution to aesthetic judgments, the more our aesthetic experiences are opaque.

In Chap. 8, Patrizia Lombardo examines 'Literature, Emotions and the Possible: Hazlitt and Stendhal'. Her main contention is that literature offers knowledge through its capacity to portray human psychology, a view which, as she notes, is at odds with the main currents of literary theory with their emphasis on psychoanalytical or sociological analyses. Lombardo relies on Hazlitt and Stendhal, two authors that, like Musil, have deeply thought about literature. She begins by reminding us of philosophical scepticism relative to literature's contribution of knowledge: for instance, Lamarque gives two arguments against the idea, the argument that literature does not provide data and the argument of banality to the effect that we already know the general truths that literature convey. Currie is ready to attribute to literature, because of the relation between fiction and emotions, some link with psychological truths. But other contemporary philosophers, such as Nussbaum or de Sousa, have different views. De Sousa, developing a Musilian argument, proposes that literature represents potentialities, actual dispositions of particular persons or things. Finally, literature enhances empathy. Lombardo insists that literature is not the way to action but rather a basis for reflection, because it develops the imagination, quoting Caroll's views on the parallelism between fiction and thought experiments, as a means to imagining the possible. According to Musil, possibilities belong to reality and literature awakens us to general possibilities via particular cases. Lombardo proposes that the distinction appropriate for literature is not particular/general, but rather contingent/essential. Literature makes available essential truths, but escapes banality by the vividness with which it portrays contingent events. It is in the balance between essential and contingent truths that style provides that literature finds its value, as Musil, Hazlitt and Stendhal insisted. The last two were similar in their position relative to the Romantic movement, in their interest in the affective, their opposition to sentimentality and their suspicion of Jean-Jacques Rousseau. Both were interested in the complexity of feelings and emotions and in the corresponding nuanced values. In that, as in their views on literature, these two eighteenth-century authors are very near to Musil.

In Chap. 9, 'L'avenir du crétinisme', Pascal Engel addresses stupidity or foolishness as insensitivity or blindness to cognitive values. Engel notes that foolishness is commonly considered as a cognitive or intellectual deficit, though this can be interpreted either in terms of *competence* (as a lack of the faculty of judgment), or in terms of *performance* (as the inability to correctly apply one's faculty of judgment). Contemporary psychology of reasoning has followed this path, leading to the conclusion that human reasoning suffers from major competence problems. This conclusion has been contested on several bases, of which the most important is that the deficits noted are in fact performance rather than competence deficits, due to the highly abstract experimental paradigms used. People use fast and frugal heuristics, ecologically adapted to the normal circumstances of human reasoning. In any case, foolishness is given an intellectual definition, which is far from previous views, such as 'classical foolishness' or 'romantic foolishness'. Both differ from the contemporary view in seeing foolishness not as a failure in judgment, but as a failure in reason, where reason includes both aesthetic and moral values. The main distinction between them is that where classical foolishness sees foolishness as a negative reverse of reason, romanticism is fascinated by it. Foolishness is a lack of sensibility, of respect for the values of reason, such as truth. In romantic foolishness, the values of the true are also denied, but the romantic attitude is a mixture of despise and fascination, through the affinity between the aesthetic values of the sublime and foolishness. Additionally, foolishness is the province of the mob, the attribute of everyone, and these two views meet in the idea that foolishness is the embodiment of reasons through tautologies. Though foolishness is multifaceted, making it difficult to choose between different accounts, Engel favours the classical view, which was also that defended by Musil. Engel notes the parallelism between that view of foolishness and the view defended by Frankfurt on 'bullshit', that it is a total disregard for, or neglect of, truth. As Engel notes, this is not incompatible with intelligence, including academic intelligence, as was shown by Sokal and Bricmont. Finally, Engel wonders on the reason why foolishness is so widespread in philosophical academic circles. He traces it first on the media exposure of philosophy in France and continental Europe, but extends this view to the possibility that the values of journalism have invaded academia everywhere, producing a new and generalized form of snobbery.

1.2 Part II: Epistemology, Perception and Consciousness

In Chap. 10, Diego Marconi discusses 'Three Easy Points on Relative Truths'. He defines relativistic semantics as including not only possible worlds but also perspectives in the truth-value assessment of propositions. Marconi notes that, though *absolute* relativism, according to which every proposition should be assessed in this way, is no longer considered tenable, *moderate* relativism, according to which only some propositions (e.g. expressing issues of taste) should be, is more often accepted. The inclusion of perspective allows for *faultless disagreement,* in which two

contradictory propositions, asserted by different speakers, can nevertheless both be considered true. In other words, moderate relativism is valid in case there are genuine cases of faultless disagreements and not all cases of disagreement are faultless. Even though moderate relativism seems reasonable, there is a problem with the notion of a relative truth, or, more precisely, with the notion that something is *true for someone*. What is the meaning of *true* in the expression *true for someone*? Either it means the same thing as *true* (*simpliciter*), and then it is not clear what *for X* adds to the notion of true. Or it means something else and we need an account of what it means exactly. Marconi compares truth depending on a context and truth depending on a perspective and notes that while the first has to do with inferential potential, the second has to do with truth-value evaluation. It is hard to know exactly how it works. A possibility is to consider subjective worlds, i.e. possible worlds distorted by someone's standards and values. This, however, would utterly destroy the notion of faultless disagreement by making disagreement impossible. A second problem has to do with the ontological status of subjective worlds: They would seem to imply subjective facts and those are strange indeed. Marconi then turns to the relation between subjective truths and beliefs. Marconi notes that relative truth cannot be simply considered on a par with belief for even matters of taste could be 'natural' (depending on physical objective relations) and thus in contradiction with the speaker's beliefs. A second problem is that relative truth cannot be a subspecies of simple truth, but has to be autonomous from it. Otherwise, (again) faultless disagreement would be impossible. In other words, a sincere assertion, though it entails the speaker's belief in the truth of the proposition asserted, does not entail that the proposition is indeed true for the speaker (relatively true). Moderate relativism implies that for each proposition it is either the case that it is simply true (or simply false), or true (or false) for someone one. But an alternative view is to consider relative truth as basic and to define objective truth as *true for everyone*. The question thus is whether objective truth, so defined, is tantamount to simple truth. There is, however, nothing to prevent a matter of taste to be universal in an accidental way. The notion of objectivity needs something stronger however: It should mandate a priori intersubjective agreement. This is tantamount to reinstalling a distinction among propositions, between those that are simply true (or false) and those that are true or false for someone. Thus, the notion *true for X* is still in need of an account.

In Chap. 11, 'Mere Belief as Modification', Maria van der Schaar is concerned with the relation between truth and belief, which she examines from the vintage point of linguistic phenomenology. Her point of departure is Williamson's contention that belief is a kind of *botched* knowing. 'Botched' is a modifying term, on a par with 'false' or 'fake', and as such is nonattributive (e.g. a fake Rembrandt is not a Rembrandt). Van der Schaar investigates the relation between belief and knowledge. A first possibility is to explain knowledge from belief. The only meanings of belief relevant here are 'capacity to judge' and 'conviction'. Both seem indeed involved in the explanation of knowledge, but though both are mental, the first seems rather linked to an act while the second refers to a state of mind. This duality goes well with the fact that knowledge is both the capacity to know that, e.g. Prince William is a father, and the act of judging that actualizes this capacity. There

are two possibilities: Either knowing depends on believing, which in turn depends on judging; or both knowledge and belief depend on judging. An argument for the second view and for an additional independence between knowledge and belief is that 'believe' is neg-raising ('John doesn't believe that *p*' is ambiguous between 'It is not the case that John believes that *p*' and 'John believes that not-*p*') while 'know' is not. Hume and Locke also advocated a radical separation between knowledge and belief. Additionally, knowledge and belief give rise to different types of question. Though it is natural to ask *why* (rather than *how*) someone believes that it rains, it is natural to ask *how* (rather than *why*) she knows that it rains. In other words, knowledge is questioned as a kind of act of coming to know while this is not the case for belief. On the other hand, beliefs are treated as mental states, resulting from a cause. A third possibility is to explain belief from knowledge and it is here, given that belief can be erroneous or *misfire,* that Williamson's notion of botched knowledge comes back. Just as a misfiring speech act is a botched speech act, *mere* belief or *mere* judgment can misfire, and 'mere', as 'botched' does not enrich the term 'belief'. This does not imply that belief is not involved. A 'mere' belief is a belief that fails to attain knowledge. Mere beliefs are not a subspecies of belief, however, as any belief can turn out to be of that sort. The word 'mere' is syncategorematic, functioning like a *restrictive* operator on the term that follows. It is not attributive, and means something like 'nothing more than' while 'pure' means something like 'nothing less than'. Van der Schaar proposes a classification of nonattributive terms, inspired by Twardowski, and concludes that 'mere' or 'pure', though nonattributive, are not modifying. She then turns to linguistic phenomenology and notes that both from a first- and from a third-person perspective, assertion or judgment and knowledge are conceptually related. The term 'mere', however, introduces a wedge between the first- and the third-person perspective: Something may seem to be a judgment from the first-person perspective and be a *mere* judgment from the third-person perspective, because it is not knowledge or true judgment. The type of modification at play is not so much a semantic as it is a conceptual modification.

In Chap. 12, Fabrice Teroni discusses 'The Epistemological Disunity of Memory'. While the intuition that memory is trustworthy is strong and consensual, the justification for memory judgments is hotly debated. Two views have been proposed: In the Past Reason Theory, what justifies our present judgments about the past is what justified them at the time; in the Present Reason Theory, what justifies them are present reasons. Teroni begins by introducing the distinction between *semantic* (or *propositional*) memories, where the verb 'to remember' typically takes a sentential complement, and *episodic* memory, where the complement will typically be nominal. An additional way of distinguishing between the two is to say that episodic memory includes, while propositional memory does not, information about when and where the event occurred. Beginning with propositional memory, Teroni addresses the questions of justification. According to the Present Reason Theory, one may appeal either to memory impressions, inference or the subject's trust in the source of the judgment. Memory impressions do not fare well for propositional memory, because it is not clear that there is always any distinct phenomenology in propositional memory judgment over and above the judgment itself and when there

is it seems uniform and thus independent of the content being remembered. The second possibility is that a propositional memory judgment is justified if it is the result of a justified inference. Despite its appeal, this suggestion is problematic: first, it appears that children do not make such inferences before they are 5 years old while we probably would not want to claim that they are incapable of propositional memory before that age; second, we do not usually seem to rely on inference in making propositional memory judgments. The third possibility is source memory, i.e. backing our propositional memory judgments by appealing to the source from which we drew the information. This, however, is again problematic, first because it is not psychologically realistic, as we usually cannot identify the source, second because it passes the epistemological burden on without solving it, as the problem becomes the justification of the source itself. The Present Reason Theory thus does not seem to offer much hope for the justification of propositional memory judgments and Teroni traces this difficulty to the fact that it relies on present-tense internalism in the sense that the justification is dependent on psychological accessibility and that this accessibility has to be present when the judgment is made. Teroni thus proposes to look at the Past Reason Theory and notes that propositional memory judgments are justified when a reason justified the *belief at the time it was accepted,* as well as at the time the memory report is issued, even if the reason cannot be accessed anymore. Thus, the Past Reason Theory is far better to account for the justification of propositional memory judgments. What about episodic memory judgments? According to Teroni, the remembering is highly different in propositional and in episodic memory: In episodic remembering, there are more than memory impressions. There are memory experiences and these experiences are highly dependent on the content being represented. Additionally, my present memory experiences represent what my initial experiences presented and thus inherit their intentionality. These experiences accompanying episodic remembering are the justification for the associated episodic memory judgments and this goes well with the Present Reason Theory. This, however, seems to sever the justification of episodic memory judgments from past reasons. But, while this would be a valid reason to reject the application of the Present Reason Theory to the justification of propositional memory judgments, it is not a good reason to reject its application to episodic memory judgments. The justification yielded by memory experiences is due to the fact that these experiences depend on the initial experiences the subject underwent when the remembered episode occurred. Thus, episodic and propositional memories do not have the same epistemology.

In Chap. 13, Göran Sundholm describes ‘The Vocabulary of Epistemology, with Observations on Some Surprising Shortcomings of the English Language’. Comparing different European languages, Sundholm notes that some distinctions lexically expressed in a number of languages (e.g. the distinction between knowing an object and knowing a truth: *connaître/savoir* or *kennen/wissen*) are not lexically expressed in English. This leads to an ambiguity in phrases such as ‘knowing a proposition’ and makes some terms, such as *Gewissheit* difficult, if not impossible, to translate. Similarly, the English word ‘evidence’ is appropriate for *evidence for* but no term is available for the notion of *evidence of*. Other examples are *proof* and *proposition*. Sundholm discusses all these terms and proposes different solutions.

Chapter 14, 'The Blurred Hen', by Clotilde Calabi, examines the problem of blurriness. Calabi begins by noting the distinction between the adverb 'blurrily', as applied to the verb 'to see', and the adjective 'fuzzy', referring to a property of what is seen. She is specifically interested in what happens when someone, looking at something with her spectacles on, non-blurrily sees it, and then, looking at the same thing without her spectacles, blurrily sees it. Calabi begins by introducing Mulligan's notion of *primitive visual certainty,* where certainty is not dependent on cognitive activity. One can distinguish, following Mulligan, two varieties of simple seeing, seeing things directly in virtue of visual content (the way we see them) and seeing some aspect or feature of what we see. The accounts of blurriness Calabi discusses rely on the first variety, involving visual content. Calabi begins with accounts according to which blurrily seeing something is not seeing it as fuzzy. According to such accounts, an *x* can look different to an observer without appearing to have changed and this is what happens in cases such as the one described above. Thus, objects that are blurrily seen are not thereby seen as fuzzy. Blurriness is a property of the visual field that either is or appears to be fuzzy. These views are subject respectively to the objection that blurrily seeing does not require an unperceived fuzzy visual field, and that what it is seen is not the visual field. Dretske sees experiences as representations, that is, as content-carrying vehicles. Blurriness can be a property of either the vehicle or the content. In the second case, it has to be represented in the content. If it is a property of the content without being a property of what is seen, then blurred perception is an illusion, a misperception. However, where is the mismatch? It might lie in the experience of the absence of details, but what is experiencing the absence of something? A third view, which does not meet with such difficulties, proposes that blurrily seeing something is seeing it not well enough to ascertain its surface details. Thus, blurriness is not an illusion but simply poor vision, leading to lack of information rather than mismatch of information. This, however, is not specific to blurred vision. Calabi proposes to keep the idea that seeing blurrily is seeing poorly, but proposes to reintroduce the notion of illusion into the account. She adopts Tye's distinction between properties *of* content (e.g. truth) and properties represent *in* content (e.g. being red). Blurriness can induce perceivers to take a property of content to be a property represented in content and this is an illusion, indeed the type of illusion commonly associated with blurriness.

In Chap. 15, Alberto Voltolini examines 'How Picture Perception Defies Cognitive Impenetrability'. Voltolini defines *cognitive impenetrability* as the impermeability of the phenomenal and intentional content of perceptual states to their subjects' cognitive states, notably beliefs. Paradigmatic examples are visual perception, as manifested in visual illusions. Cognitive impenetrability is directly linked to Fodorian *modularity*. Ambiguous figures, like the duck–rabbit, where what one believes to be represented influences what one sees, are rather obvious counterexamples to the cognitive impenetrability of visual perception. Voltolini admits that there may be some nonconceptual ways of switching from seeing the picture as a duck to seeing it as a rabbit and vice versa. But when the switch concerns the whole figure (when it is a *Gestalt switch*), concepts enter as such. The distinction might be linked to the distinction between *exogenous* and *endogenous* forms of attention,

of which only the second implies cognitive penetrability. What is more, all pictures are potentially ambiguous. Hence the same conceptual involvement occurs in those cases, as illustrated by James' well-known picture of a Dalmatian. Additionally, depending on the concepts involved, the subject will, or not, undergo a phenomenal change: Seeing the picture as a duck or as a rabbit implies different phenomenal experiences while seeing it as a rabbit or as a hare does not. Voltolini acknowledges that some Gestalt switches may be 'purely optical', rather than conceptually induced. But to contest the thesis of the cognitive impenetrability of perception, it is enough to show that, in *some* cases, concepts are involved. In fact, there are different kinds of seeing-as, some compatible with cognitive impenetrability while others are not. In other words, in some cases at least, visual perception is 'half visual experience, half thought', in Wittgenstein's terms.

In Chap. 16, Philip Gerrans deals with 'Singular Thoughts, Seeing Doubles and Delusional Misidentification'. Gerrans begins by listing the different misidentification syndromes: Capgras delusion, Fregoli delusion and the delusion of intermetamorphosis. All involve a mismatch between the perceptual representation of an individual's face and that individual. These syndromes illustrate a fundamental problem for most theories of perception: Perception is taken to concern features or properties, not objects themselves. In other words, it would track *qualitative* rather than *numerical identity*. Changes in an object will be perceptually represented as changes in its properties. The main question is not whether perception tracks objects (it does), but whether this is all that it does. In other words, does perception only allow us contact with bundles of properties or with the objects themselves? Mulligan has defended the view that they do both. According to Gerrans, there is one way of making sense of misidentification syndromes: assuming that one perceives objects, in this instance selves, rather than only their properties. In other words, when we see a face, we see a person, not a configuration of features. According to the *metaphysical* version of descriptive theories, objects are just bundles of features. This seems to be contradicted by the semantics of referential expressions. Given that perception and singular reference are linked in that the ability to refer seems to depend on a prior perceptual ability to identify individuals, though descriptive theories of perception can account for qualitative identity (as in definite descriptions), they cannot account for numerical identity (as in singular reference). The solution seems to go through a demonstrative sense, referring to a *veridical perceptual content,* which completes it. If this is successful, the content of the perception must identify an individual. This implies that the content is nonconceptual. Additionally, perception is dynamic: It tracks objects and this implies that the identity of objects is part of their perceptual representation. In other words, appearance and identity are different facets of perceptual representation, and though normally associated, can be dissociated. This is what happens in delusions of misidentification. Gerrans turns to one specific case of Capgras delusion, proposing an analysis that can be extended to other types of delusions of misidentification. Gerrans' analysis is an extension to persons of the analysis proposed above for objects, and relies on the notion of *person file*. It relies thus on four factors: appearance, familiarity, semantic information and person file. Dysfunction in any of these will produce one or another type of delusion of misidentification.

In Chap. 17, Alfredo Paternoster proposes ‘Reconstructing (Phenomenal) Consciousness’. Paternoster begins by acknowledging that the words ‘conscious’ and ‘consciousness’ can have various uses, notably the so-called *transitive* and *intransitive* uses. Paternoster is especially interested in Block’s distinction between *phenomenal* and *access* consciousness (P-consciousness and A-consciousness), P-consciousness being essentially first personal. Paternoster rejects this dual view of consciousness and proposes a *unique* notion: Consciousness is fundamentally phenomenal, but phenomenal effects presuppose access. He begins by criticizing Block’s distinction on several counts. First, it is applied to mental states while consciousness is a property of subjects. Second, it is because of the distinction that P-consciousness is seen as precluding any scientific account. The main support for Block’s distinction comes from cases of double dissociation, where our intuitions are wavering, and which can be explained through the distinction. P-consciousness without A-consciousness is taken to occur when one suddenly gets access to a conscious experiential state that was present, but not accessible before (when our attention is otherwise engaged, for instance). However, Crane has proposed an alternative, more convincing, account, according to which this is simply a case of inattentional consciousness. Cases of A-consciousness without P-consciousness are, on the other hand, hard to find, the only suitable example being the (fictitious) *super-blindsight*. In regular blindsight, indeed, the visual information is *not* freely accessible, as it should be in A-consciousness. This should not be surprising as it seems counterintuitive to attribute consciousness to a subject when she has no phenomenal consciousness. Paternoster proposes the view that conscious contents can be more or less accessible, depending on, e.g. attention. An alternative possibility is to resort to the notion of nonconceptual content, but this is unsatisfying for several reasons. A better approach would postulate degrees or levels of accessibility.

Martine Nida-Rümelin discusses ‘Basic Intentionality, Primitive Awareness and Awareness of Oneself’ in Chap. 18. Nida-Rümelin begins by noting that questions of consciousness are more and more couched in first-personal terms (*what it is like for **me***). In other words, a new factor, a mine-ness or a for-me-ness is introduced into the notion of experience. Nida-Rümelin claims that this leads to a number of mistakes. There are, indeed, three different aspects of consciousness that *subjective character* can be taken to refer to. The first, basic intentionality, has to do with the fact that ‘in any experience there is a subject *to whom* something is phenomenally given’. The experience has *basic intentionality* in the sense that it is possible to distinguish between the experience itself (an event involving a subject) and the experienced (what is phenomenally *given* to that subject). The second aspect is primitive awareness: the subject not only has a given experience, but is aware of having that experience. This awareness is constitutive of the experience, but it does not involve a conceptual judgment as to the nature of the experience. Being primitively aware of one’s experience means that the experience is partly constitutive of one’s phenomenology. The third aspect is awareness of basic intentionality. It has to do with the question of whether one must be aware of oneself in every experience. According to Nida-Rümelin, the answer is positive. Given that an experience exhibits basic intentionality, this means that the subject is aware of the structure basic

intentionality imposes on experience. This does not, however, require a concept of oneself, nor does it require that the subject *experiences* basic intentionality (this would lead to a regress). In other words, basic intentionality does not add anything to the content of the experience. The notion of *subjective character* is used ambiguously to refer either to basic intentionality, primitive awareness or awareness of basic intentionality. The question of whether subjective character is part of experience should accordingly be divided in three. If subjective character is understood in terms of awareness of basic intentionality, the answer is positive. If it is understood either as basic intentionality or as primitive awareness, the answer is negative. Finally Nida-Rümelin discusses and rejects the notion of peripheral inner awareness (proposed by Kriegel), as well as self-representationalism and representationalism. In conclusion, she uses the three notions described above to debunk some erroneous views linked to the notion of subjective character. In an appendix, she quotes and comments various passages relative to the notion of subjective character.

1.3 Part III: Philosophy of Mind and Philosophy of Language

Part III opens with Uwe Meixner's discussion of 'Causal Equivalence as a Basis for the Specification of Neural Correlates', in Chap. 19. Meixner begins by defining *causal equivalence* between two events as both events sharing all their causes and all their consequences. He then applies the notion of causal equivalence to the relation between mental and physical events, leading to a strong possibility—every mental event is causally equivalent to some *physical* event—or to a weaker possibility—every mental event is causally equivalent to some at least *partly physical* event. There is, of course, the possibility of causal nonequivalence, with again, a strong thesis—for some mental event, there is no partly physical causally equivalent event—and a weaker thesis—for some mental event, there is no wholly physical causally equivalent event. From that standpoint, the goal of cognitive neuroscience should be to adjudicate between the causal equivalence and the causal nonequivalence theses. As Meixner notes, causal equivalence not being identity, establishing causal equivalence would not thereby contradict dualism. The ongoing quest for mental correlates can be interpreted as the quest for physical events causally equivalent with mental events. This has a few implications: physical events do not cause the mental events they are causally equivalent to; mental events are not epiphenomenal relative to wholly physical events; purely mental events with neural correlates give rise to causal overdetermination; every mental event with a neural correlate is identical to that mental correlate because of physical causal closure, leading to an insolvable philosophical dilemma.

In Chap. 20, 'Simulation Versus Theory-Theory: A Plea for an Epistemological Turn', Julien Deonna and Bence Nannay discuss the controversy over the theory of mind. The authors begin by presenting the two contenders, i.e. simulation and theory-theory. As they note, there seems to be a consensus to the effect that mind reading

implies both types of processes. This, however, does not mean that the debate is obsolete. The authors' contribution is a clarification of the issues involved. Two crucial concepts in the simulation view are imagination and attribution. The question Deonna and Nannay raise is how imagination can justify (constitute a reason for) attribution. If there is need of an intermediate step, relying on a psychological theory, then any account of mind reading has to be hybrid. If no such intermediate step is needed, we can end up with a 'pure simulation' account. As the authors remark, this approach shifts the debate toward epistemology. Simulation implies imagination, and more specifically imagination *de se,* i.e. *self-imagination,* or *imagining from the inside oneself in another's situation*. This is ambiguous between imagining oneself into a physical position (actually occupied by the target) or imagining oneself as having the target's psychological make-up. Both interpretations are possible given the rather disparate goals of simulation, which may or may not entail the most costly second possibility. In any case, imagining oneself in another's situation basically means imagining oneself in what one *takes to be* the other's position. Finally, imagining can be or not be intended. Attribution can be simply described as representing the other to have a certain mental state. Thus, simulation works as follows: I imagine myself as in your situation, and, *based on this imagining,* I attribute a mental state to you. The shape of the debate between simulation and theory-theory depends on how one interprets the *reason* that imagining is supposed to be for attribution. The crucial question is whether the reason or justification in question presupposes the application of a psychological theory. If it does, then simulation necessarily involves theorizing and a hybrid model is mandatory. If it does not, then we end up with pure simulation, and we have to account for how imagination *justifies* attribution. Deonna and Nannay make an analogy with reliabilist theories of perception. One might suppose that the same kind of relation between imagination and attribution obtains. In this case, pure simulation is possible.

In Chap. 21, Jérôme Dokic pursues the subject, dealing with 'Mental Simulation and the Reification of Beliefs'. Dokic's aim is to clarify the respective contributions of simulation and theory in a hybrid mind-reading system. Dokic begins by outlining simulation theory, whose main tenet is that imagining or simulating is a way of acceding to others' mental states. However, this seems far short of ascribing a belief. Engaging in simulation is one thing, exploiting the results of the simulation another. There are two main questions here: whether belief-ascription through simulation deploys psychological concepts; and whether belief-ascription through simulation involves the concept of belief as a mental state. According to Dokic, a nonconceptual view is impossible, which means that belief-ascription has to deploy psychological concepts. Theory-theory would insist that it also needs the concept of belief as a mental state while simulation theory denies it. Dokic compares mental simulation and make-believe games. In make-believe games, the doxastic states of the agent and her representation of the imaginary world do not conflict because they belong to different mental models. Engaging in make-believe does not entail a representation of the pretence as such. However, a reflective creature will embed the representation of the imaginary world into her doxastic representation. There is no reason to think that it should be any different in mental simulation: An imagined

situation can be indexed to the specific individual whose mental states are being simulated. Dokic proposes to consider '*X* believes that', on a par with 'according to *X*', as the marker of such an indexation for simulated mental states, calling this the 'quasi-modal account of belief'. On this account, belief-ascriptions deploy psychological concepts, without introducting beliefs as (abstract) objects. Gordon proposes that an appropriate embedding of simulations can produce the sophisticated notion of belief, that entails the possibility of false beliefs, leading to a potential dissociation between one's own doxastic states and the facts. The problem for the simulation account is that this cannot work. Indexing imaginary situations does not entail that the simulator apprehends them as possible world alternatives to the actual world. In other words, it does not, in and of itself, give an understanding of false belief. Dokic proposes an analogy with tense logic, where sentences are indexed relative to times, introducing an asymmetry between the present and other times. On the quasi-modal account of belief, there is a similar asymmetry between one's own doxastic states and others'. Hence, ontological neutrality is impossible. Dokic turns to what could satisfy the minimal requirement for the concept of belief. He introduces the Reification Argument, according to which it is satisfied via reference or quantification over our doxastic perspectives or beliefs: in other words, 'folk psychology reifies belief', and this agrees with theory-theory. Belief ascriptions are metarepresentational in a strong sense. Simulation, on this view, plays an epistemological role in facilitating access to others' mental states.

In Chap. 22, Roberto Casati examines 'Numerals and Word Sequences'. Casati begins by noting that, as linguistic items, numerals have atypical properties both from the syntactic and from the semantic points of view. They share some of them with expressions such as the names for days, months, fingers, letters of the alphabet and musical notes, though, by contrast with these, numerals can compose (e.g. 'one hundred and twenty seven'). What makes these features surprising is that they do not come from intrinsic properties of numbers. From a cognitive point of view, there are two systems of 'numerosity representation', the first yielding an exact representation of small quantities while the second yields an approximate representation of larger quantities. These are not triggered by the same perceptual stimuli and are mutually exclusive. They seem to be innate. What is the role of numerals relative to these two cognitive systems? A first advantage of numerals is allowing precision. This, indeed, is the line taken by Spelke and Tsivin who suggest that numerals constitute a cognitive bridge from smaller to larger quantities, allowing for precision all over the board. The bridging factor is the numeral system of natural language, linked to the fact that language has two features that innate systems of numerosity lack: no domain specificity and combinatoriality. Despite its intuitive appeal, this hypothesis meets with two difficulties: first, combinatoriality does not seem crucial to the numeral system, as quite a few numerals are not combinatorial; second, precision already exists in the system for representing small cardinalities and is only necessary in the system for representing larger cardinalities when they are very similar (can only be distinguished by counting). Thus, precision comes with counting rather than the representation of cardinality as such. Casati then turns to the acquisition of numerals and notes that they are learnt as part of a sequence and independently of

the semantic meaning of each component, rather as one would learn a rhyme, and indeed independently of the innate systems for numerosity representation. Casati also notes that this strongly distinguishes learning a word such as 'two' and learning a word such as 'dog'. According to him, this mirrors a morphological difference: Numerals are (literally) part of the sequence of numerals while 'dog' is not part of any specific sequence. This leads Casati to his second hypothesis, according to which the sequence of numerals is a map where numerals are situated relative to one another, and where their semantics can be accessed through their positions, once the semantics for small numerals is in place. In other words, numerals, as well as other sequential lexical items (days, months, etc.) enumerated above, constitute a small 'artificial language' appended to natural language. This view is compatible with a number of hypotheses relative to the origin of numerals, notably with the view that they stemmed from the use of fingers for counting. Thus, numerals are not a bridge between the representation of small and large cardinalities. They are just the next step in number cognition after the numerosity representation system specific to small cardinalities.

In Chap. 23, Jonathan Barnes describes and comments on 'Frege's New Language'. Barnes begins by noting that philosophers have always lamented the shortcomings of natural language. Frege invented a new language, a *Begriffschrift,* an ideography, to solve the problem. Fregean—Frege's *Begriffschrift*—was adequate for arithmetic and logic in the sense that it could express without unnecessary adornments and perspicuously any arithmetic or logical thought. Frege's first criticism of natural language was structural ambiguity, as manifested in definite descriptions, which can be interpreted as referring to a specific individual, genera, group, etc. In Fregean, there is a sign which uniquely represents the specific interpretation. But why should the polyvalence of the definite description worry us? Speakers usually can easily and successfully interpret definite descriptions. Where Fregean has an advantage over natural language is in making clear the referential potential of thoughts which natural language hides. This, however, ignores the distinction between surface and deep grammar: Structures that seem identical at the surface level have contrasted structures at the deep level. Frege also complained that inferential adverbs in natural language—e.g. 'therefore', 'so'—are promiscuous (not restricted to valid inferences). On the other hand, Fregean is restricted to the expression of only one form of (valid) inference, the modus ponens. One major difference between English and Fregean is that Fregean can express thoughts far more complex than can English that either cannot express them at all, or can only express them in an unintelligible way. However, this is hardly a problem for English: If no competent speaker of English can understand them, then it is a moot question whether they need to be expressed at all. It is true that some scientific thoughts can only be interpreted in Fregean and not in natural language, but expressing them is hardly a common necessity.

In Chap. 24, Wolfgang Künne speaks 'On Liars, "Liars" and Harmless Self-Reference'. Künne first recalls the distinction, borrowed by Bolzano from Aristotle, between fallacies due to verbal expression and fallacies that are not. An example of the second occurs when something said with qualifications is understood without

qualification (e.g. when something is said to be both *S* and not-*S*). Another possibility is when someone says simultaneously something true and something false. Bolzano turns to what Künne calls the *argument of the Self-Confessed Liar*. The Self-Confessed Liar is complicated by the fact that being a liar depends on insincerity rather than on saying what is false. Insincerity occurs when the speaker says something she believes to be false. Thus, lies differ from deception in that they do not necessarily entail saying something false, though liars are often taken to try to deceive their addressees. In Bolzano's view, a liar has two deceptive intentions, a *thematic* one, in which the speaker intends her hearer to acquire a false belief relative to the topic of the proposition expressed; an *attitudinal* one, according to which the speaker intends her hearer to think that she is sincere, in saying what she says. However, Bolzano's account, though appropriate for lying with deceptive intent, seems too narrow: Some lies may not involve a thematic deceptive intention. Another complication is added when one considers not only what is explicitly said, but also what is implicitly communicated. However, lies should be limited to what is explicitly asserted content: in other words, insincerity is not sufficient for an utterance to be a lie. Bolzano mistook Eubulides' Pseudónemon for the Self-Confessed Liar. The Self-Confessed Liar ('I am a liar') is obviously harmless from a logical point of view, something that Bolzano recognized. There are contemporary versions of what has come to be called 'The Liar' that are not, however. The first mention is found in Cicero and subsequent mentions begin with Russell, who translated the Latin and Greek sentences as 'I am lying', ignoring the ambiguousness of the verbs, which can mean either to lie or to say something false. Both 'I am lying' and 'I am saying something false' are different from the Self-Confessed Liar ('I am a liar'), and, additionally, only the first has to do with lying, as seen above. It is, however, the second translation that should be preferred, and Künne proposes to call it the *antinomy of falsity* or *F-antinomy,* rather than 'The Liar'. A *paradox* is a statement that seems obviously false drawn from obviously true premises. An *antinomy* is a self-contradictory statement. Is the Russellian version of 'The Liar' ('I am lying') truly an antinomy? Künne answers in the negative, as lying does not consist in saying something false. To get a paradox, we need a reflexive reading of the F-antinomy. Künne then turns to Epimenides the Cretan and his pronouncement that all Cretans are liars. Russell borrows it from an alleged letter by Paul the Apostle. Ever since, reference to Epimenides has become a commonplace in the literature on the F-antinomy. Russell's analysis of Epimenides' sentence is faulty, however, because the sentence can be a lie and true; hence, no paradox arises. Though what Epimenides did say is not paradoxical, some alternative versions (e.g. 'Whatever is asserted by a Cretan is false') are and some are antinomies (e.g. 'Epimenides the Cretan said that all Cretans are liars, and all other statements made by Cretans are certainly lies'). Künne then turns to Savonarola's treatment of the F-antinomy. Savonarola used an explicitly self-referential version ('*This* is false', where 'this' is reflexive) and classed it among paradoxes. As Savonarola refused the notion that a sentence can fall into a truth-value gap, he claimed that 'This is false' was not a sentence. Though Savonarola is not the only philosopher to adopt this position, it seems difficult to endorse it. First, bivalence is not beyond doubt. Second, the argument

does not stand. Third, some sentences can appear paradoxical in some empirical situations. Finally, the idea that such sentences, while linguistically well formed, are not sentences does not make sense, as noted by Bolzano. Bolzano's proposal is that such sentences are false, and, again, this is a position shared by other philosophers. Nevertheless, it is not quite right, and a better description of such sentences is that they do not seem to have a consistent, *grounded,* truth-value. Künne then turns to self-referentiality and notes that it is not necessary to trigger antinomies: Reciprocal reference is enough. It is not sufficient either, as all self-referential sentences are not ipso facto antinomic. This is not to say, however, that self-referential sentences do not have remarkable properties, already noted by Bolzano. In a series of appendices, Künne comments on various translations, versions and accounts of 'the Liar'.

In Chap. 25, Manuel García-Carpintero discusses 'Constitute Versus Normative Accounts of Speech and Mental Acts'. Mulligan and others have defended the view that constitutive accounts of meaning, being based on internal relations, cannot also be normative. The later Wittgenstein argued that there was a constitutive connection between meaning and normativity. Speech acts have representational contents, i.e. propositions, which encode correctness conditions. These, however, are not normative reasons. There are, of course, other rules relative to speech acts and, notably, to assertion, though, on a Gricean view, these rules are regulative, rather than normative. García-Carpintero, however, notes that nonnormative accounts are subject to the objections against Gricean accounts and that, intuitively, we have the intuition that rules governing assertion are normative, rather than merely regulative. Thus, speech acts are normative, even when they are implicit, as in conversational implicatures. García-Carpintero then proposes a generalisation of Kaplan's distinction between character and content. Natural language semantics is 'character'-semantics. Semantic normativity, however, is weak. What is needed to account for the normativity of speech acts are rules that constrain the agent's activity, something more than a simple Lewisian convention, something like a social contract.

In Chap. 26, Laurent Cesalli examines 'M&Ms—Mentally Mediated Meanings'. Marty repeatedly mentioned the scholastic claim that 'words signify things by means of concepts'. In consequence, according to Marty, strict synonymy occurs when different words designate the same object via the same concept; nomination depends on indication; proper names are mediated by singular concepts; for nouns with conceptual presentation (e.g. white), the mediation is done via the concept content; proper and improper presentations differ in their mediating concepts, improper presentations being non-perceptual. On Marty's analysis, the name is used to indicate a certain concept in the speaker, which is triggered in the hearer and this presentation is intentional. Marty was mainly relying on Boethius' commentary of Aristotle. Boethius identified a double dimension of cognition and communication, which led to two divergent but compatible views: a semantic one and a pragmatic-semantic one. Cesalli reviews some medieval interpretations of the claim, Abelard's, Aquinas', Bacon's, Ockham's and Buridan's. Though Marty's position differs from medieval positions from a strictly semantic point of view, it is very similar to them from a pragmatic-semantic point of view, and it is very near to Bacon's position, with its emphasis on communicative interaction.

In Chap. 27, François Recanati writes about 'Mental Files and Identity'. Recanati begins by introducing Frege's modes of presentation or senses in relation to opacity and informativity of identity statements. Though Frege saw modes of presentations as essentially descriptive, this meets with a number of difficulties. Recanati proposes to see them nondescriptively as mental files, i.e. as individual or singular concepts. The (referential) relations on which mental files are based are *epistemically rewarding* in that they allow the subject to gain information over their objects. The information in one mental file is insulated from the information in other mental files. This limitation is overcome when the subject makes a judgment of identity whereby two different files actually refer to the same individual. This *linking* is not merging: The two files remain different, though information can flow between them. By contrast with judgments of identity, presumptions of identity are operative inside a single file and, indeed, allow information to accumulate about what is considered as one and a single individual. Files are strongly tied to their epistemically rewarding relations and they are thus quite often temporary, existing only as long as the epistemically rewarding relation that triggered them is in force. In some cases, the information is transferred to another more permanent file. In other cases, this is not the case, and, when this happens, they are proto-files, lacking generality, rather than *bona fide* files. While proto-files contain information gathered through epistemically rewarding relations, regular files are fully conceptual and can also contain information acquired through third persons. There are in fact two kinds of conceptual files: those based on first-order epistemically rewarding relations, and those based on higher-order epistemically rewarding relations, encyclopedic files. Recanati then answers several objections against the mental files account. The first one is the circularity objection. The account is circular because files rest on a presumption of identity, tied to de jure coreference. However, or so the objection goes, de jure coreference depends on judgments of identity, themselves tied to de facto coreference, which are accounted for in terms of operations on mental files, hence the circularity. To avoid the regress, Recanati proposes to use the hierarchy of files introduced before, with (nonconceptual) proto-files at the bottom, (conceptual) individual files in the middle and (conceptual, abstract) encyclopedic files at the top. The idea is to introduce at the level of proto-files a proto-linking operation that does not rest on identity judgments and is a first step for proto-files to attain a conceptual status. The second objection is the transitivity objection which says that de jure coreference, which rests on identity, is not transitive while identity is. Recanati distinguishes two versions of the mental files account, a strong one, where being associated with a single mental file is *necessary* for de jure coreference, and a weak one, where it is only *sufficient*. The transitivity objection is an objection against the strong one, but fails against the weak one.

In Chap. 28, Marco Santambrogio asks 'Did *Madagascar* Undergo a Change in Referent?' He begins by reminding the reader that Mulligan defends the view that 'perception and behaviour are part of language'. Indeed, perception fixes reference. In other words, there is a dependence between a singular term and its referent and Mulligan interprets it in modal terms as necessary. Santambrogio defends the view that directly referring names are indeed object dependent, but distinguishes object

dependence from Kripkean rigidity. He explores it through one specific problem: Whether a name can change its referent, taking Evans' example of *Madagascar*. Santambrogio begins by examining what happens when we use an existing name (e.g. *Napoleon*) for a new object (e.g. one's pet aardvark). In such a case, a new name is created. Santambrogio then turns to cases when names change their referents without a baptism, but through social change and linguistic evolution. Kripke's notion of a causal chain is clearly social, but static nevertheless. Kripke distinguishes the *speaker's referent* (what object the speaker refers to, whether or not the name used refers to it in the speaker's idiolect) and the *semantic referent* (the referent of the name in the speaker's idiolect). The speaker's reference can differ from semantic reference through the speaker's error. In the case of *Madagascar,* the name was originally used by Africans to refer to a region of the continent, before Marco Polo mistakenly used it to refer to the island, a usage adopted by Europeans. In this case, it seems clear that there are two different names. The interesting question is when the change occurred.

Chapter 29, by Anne Reboul, is dedicated to 'Live Metaphors'. Reboul reminds us of the distinction between *continuous* accounts of metaphors (metaphors are interpreted on a par with nonmetaphoric utterances) and *discontinuous* accounts (metaphors undergo a specific interpretation process). Typically, discontinuous accounts rely on the notion of *figurative meaning,* usually postulating that the specific process is triggered by the obvious—if not necessary—falsity of the metaphor. An obvious objection against the discontinuous view is that metaphors are not always false. What is more, negating false metaphors, though it makes them true, does not make them any less metaphorical. Finally, figurative meaning is usually a nonmetaphorical paraphrase of the metaphorical utterance and, as is well known, metaphors do not lend them to paraphrase without loss. Relevance Theory has produced two successive, and currently trendy, continuous accounts of metaphor. On the first Relevance-Theoretic account, metaphors are instances of vague communication and are to be interpreted through the production of weak implicatures. On that account, the reason why metaphors cannot be paraphrased without loss is because they weakly communicate a wide array of implicatures, which could not have been communicated as economically through a nonmetaphoric utterance. On the new Relevance-Theoretic account, metaphors are interpreted through the production of an explicature, not of implicatures, by broadening and/or strengthening one of the concepts in the logical form of the utterance, this concept being replaced by an ad hoc concept. The account is continuous because this is also the way vague utterances are interpreted. It is not clear whether it preserves the advantages of the first Relevance-Theoretic account, i.e. its ability to account for why metaphors cannot be paraphrased without loss and why they are produced in the first place. In addition, it is not clear that the notion of ad hoc concept is really compatible with basic tenets of Relevance Theory. Indeed, Relevance Theory rests on a Fodorian account of concepts according to which concepts are atomic, hence not definitions. Ad hoc concepts, however, are supposed to be formed by modifying the *definition* of the original concept by deleting features or introducing them in the definition. This directly contradicts a Fodorian view of concepts. Additionally, it is not clear that the

second Relevance-Theoretic account does not reintroduce the notion of *figurative meaning*: What, after all, is the explicature, if not a form of figurative meaning? Both Relevance-Theoretic accounts fail to account for the non-propositional effects of the metaphors and though such effects are not specific to metaphors (they are also to be found in, e.g. Japanese haikus), they certainly should be central to any account of metaphors. Indeed, metaphors produce both propositional effects (implicatures) and non-propositional (sensory) effects and thus the first Relevance-Theoretic account tells only part of the story.

In Chap. 30, Luigi Rizzi examines 'Syntactic Cartography and the Syntactization of Scope-Discourse Semantics'. Cartographic representations are at the interface between syntax and interpretation. Interpretation of linguistic expressions concerns two main properties, argumental semantics, and properties of scope-discourse semantics. Sentences sharing the same argument structure may have highly different informational (scope-discourse) properties, as happens in cleft sentences. Thematic assignment (argumental semantics) rests on the local configuration produced by the operation Merge. Informational assignments (scope-discourse semantics), on the other hand, more often than not, are due to movement, displacing elements from their thematic positions. An important question is how scope-discourse properties (focus, topic, etc.) are attributed from the position of the displaced elements. Rizzi describes the criterial approach under which assignment of scope-discourse properties is structurally based. The view postulates functional heads (topic, focus, relative, question, exclamative) with a dual function, both internal to syntax and relevant for sound/meaning interfaces. These functional heads are expressed overtly in some languages. On alternative, non-cartographic approaches, the interpretive work relative to scope-discourse semantics has to be done post-syntactically and is thus both less transparent and more complex.

1.4 Conclusion

These two volumes of philosophical essays dedicated to Kevin Mulligan give a good picture of his many interests, going from metaphysics, values, history of philosophy, all the way to ethics, epistemology and philosophy of mind and language.

Part I
Values, Ethics, and Emotions

Chapter 2
Alternatives and Responsibility: An Asymmetrical Approach

Carlos J. Moya

Abstract In this chapter, I defend an asymmetrical view concerning the relationship between alternative possibilities (APs) and moral responsibility (MR), according to which APs are required for being blameworthy, but not praiseworthy, for what one decides or does. I defend the nonnecessity of alternatives for praiseworthiness through an examination of what I call "Luther" examples. My defense of the necessity of alternatives for blameworthiness proceeds instead through an analysis of so-called Frankfurt examples. In both cases, my arguments rest on the contention that, in ascriptions of MR, the primary question is not whether the agent could have done otherwise, but whether she should have done what she did, so that the former question only becomes pressing when the answer to the latter is negative. Concerning MR, then, the concept of moral obligation or duty is prior to that of APs.

Keywords Principle of alternative possibilities · Moral responsibility · Moral obligation · Blameworthiness · Praiseworthiness

2.1 Introduction: Moral Responsibility, Alternatives and Moral Obligation

The principle of alternative possibilities (PAP) can be stated thus: "A person is morally responsible for something she has done only if she was able to do otherwise." For the last 40 years or so, this principle has been the object of a long and intricate discussion. Philosophically speaking, the really important question is not PAP itself, but the core intuition behind it, namely, that alternative possibilities (APs) are necessary for moral responsibility (MR). In what follows, with "PAP" I shall refer to this core intuition. According to some thinkers (Frankfurt 1969; Fischer 1994; Pereboom 2001, 2009), PAP is false. Others contend that it is true (Naylor 1984; Wyma 1997; Otsuka 1998). These are the two main positions about this important issue. There are, however, a minority of philosophers (Wolf 1990; Nelkin 2008) who hold an asymmetrical view according to which PAP is true for blameworthy

C. J. Moya (✉)
University of Valencia, Valencia, Spain
e-mail: carlos.moya@uv.es

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_2,

actions, but false for praiseworthy ones. In other words, APs are required for negative MR (blameworthiness, reprehensibility), but not for positive MR (praiseworthiness, laudability). In this chapter, and unlike what I used to hold, I will side with this minority position and defend an asymmetrical view of this kind.

Susan Wolf's defense of asymmetry is based on a theoretical approach to MR which she labels the "Reason View" according to which, in her own words, "an individual is responsible if and only if she is able to form her actions on the basis of her values *and* she is able to form her values on the basis of what is True and Good" (Wolf 1990, p. 75). In everyday terms, what is necessary and sufficient for MR is to be able to distinguish between good and evil and to act on the basis of the former. Now, asymmetry follows from this account. If a person behaves rightly and on the basis of correct values, this shows that she has the required ability, and so is responsible for what she did, no matter whether she could have behaved in a different and less admirable manner. Instead, if a person does a wrong thing or does a right thing but on the basis of wrong or distorted values, it is not clear whether she has the required ability and so whether she is morally responsible. For her to be so, she must be able to act in a morally right way on the basis of correct values: She must be able to do otherwise. APs are, then, required in order to be blameworthy, but not praiseworthy, for what one does. In a recent paper, and following Wolf's steps, Dana Nelkin defends a similar asymmetrical view, which she labels "the rational abilities view":

> I defend a view according to which one is responsible for one's actions to the extent that one has the ability to do the right thing for the right reasons. This view is asymmetrical in requiring the ability to do otherwise when one acts badly or for bad reasons, but no such ability in cases in which one acts well for good ones. (Nelkin 2008, p. 497)

Though not unrelated to Wolf's and Nelkin's, my own defense of an asymmetrical view concerning APs rests instead on the concept of moral obligation or moral duty. This is a central notion in our everyday conception of MR. More generally, the idea of obligation, of what one should or should not do in certain contexts and circumstances, is a crucial element in social life and social interaction. We all have strong expectations about how others (and ourselves) should behave, and react in negative ways when these expectations are frustrated. Social roles and their effective playing are central ingredients of human society. Following a relatively old trend in sociology called "symbolic interactionism," a role can be conceived as a set of mutual expectations, which involve certain obligations and, eventually, sanctions. Social institutions, in turn, would amount to organized sets of such roles. Moral expectations, moral obligations and moral sanctions are a partial aspect of this general social frame; they are an integral part of the institution of morality, with ascriptions of MR, of praise and blame, as a central aspect of it. So, far from being trivial or unimportant, the concept of moral obligation has a principal place in human life.

Now, my proposal is that we can best understand the place of the demand for APs in ascriptions of MR by going through the concept of moral obligation, of what people should or should not do or have done.

In our folk conceptual frame, I contend, MR is related to APs, to freedom to do otherwise, through the mediating notion of moral obligation. In assessing MR in

particular cases, our primary concern is not whether the agent could have done otherwise, but whether she *should* or *should not* have done what she did. And, as I will argue, the question whether the agent *could* have done otherwise is normally raised only when we judge that the agent *should not* have acted as she did. Instead, when an agent has behaved as she should, and for the right reasons, there is nothing that she should have done and did not do; this being so, there is usually no point in asking whether she *could* have done what she *should not* have done. Maybe she could; but even if she could not, this does not detract from her praiseworthiness; on the contrary, it may make her more praiseworthy than she would have been if she could have done otherwise. As Wolf has pointed out, if a person who sees someone else in danger of drowning considers seriously the possibility of not jumping in the water owing to, e.g. possible damage to her garments, we will probably consider her as less praiseworthy than another who simply does not hesitate at all in jumping in the water and does not even consider the possibility of doing otherwise. Consider also Luther's statement, when urged to retract his religious views and stop his challenge to the Roman Church: "Here I stand; I can do no other." In my 2006 book, I labelled cases of this kind, in which a person sees so clearly what her duty is that it is unthinkable for her to act in a different way, "Luther" cases. Cases of this sort have often been used to show the falsity of PAP. They play a central role in Dennett's rejection of this principle (cf. Dennett 1984, p. 133). Dennett uses in fact Luther's example, and he presents as well, as a case of this sort, his conviction that he himself would be unable to torture an innocent person for US$ 1,000. These cases play also an important role in Harry Frankfurt's work on MR; he considers some boundaries of a person's will, which he calls volitional necessities, as an integral part of that person's freedom and autonomy and a central root of MR, in opposition to a view that insists almost exclusively on the importance of choice among APs.

As is well known, and unlike Dennett, Frankfurt does not base his rejection of PAP on "Luther" cases, but mainly on cases which he himself designed and which, because of it, have come to be known as Frankfurt (or Frankfurt-style) cases. To anticipate, these are cases in which, putatively, an agent is morally responsible for what she has done though, unknown to her, she could not have done otherwise. However, in spite of these differences, it is pretty clear that both Dennett and Frankfurt take the respective cases to refute PAP concerning MR in general, that is, both as praise and as blameworthiness. This is obvious in Dennett's case, but it is also pretty clear in Frankfurt's.[1] However, in both kinds of cases we can find interesting asymmetries, related to moral obligation and the moral quality of the action.

[1] So, after presenting his putative counterexample to PAP, Frankfurt writes: "It would be quite unreasonable to excuse $Jones_4$ for his action, *or to withhold the praise to which it would normally entitle him,* on the basis of the fact that he could not have done otherwise" (Frankfurt 1969, p. 7, my emphasis).

2.2 "Luther" Cases

Let us attend to "Luther" cases first. As Dennett himself rightly points out, we do not find Luther's avowal, "I can do no other," as a sign of fragility of his rational faculties, but rather as a manifestation of their strength and soundness. And this is so even if, owing to the firmness of his convictions and values, he is actually psychologically unable to retract his religious and political stance. Concerning our moral assessment, we tend to see his avowal and the strength of his commitment as a sign of moral integrity; his inability to recant does not undermine his praiseworthiness; it rather increases it. Luther acted as he should. There is not a different way of acting that he *should* have gone for. So, even if he actually could have behaved differently, this is either irrelevant to his moral praiseworthiness or even damaging to it. Reflection on cases of this kind makes us aware that we do not demand APs in order for an agent to be praiseworthy for a morally good or admirable action. More common are examples of volitional necessities, to use Frankfurt's terms. Dennett's inability to torture an innocent person for US$ 1,000 is an example. But all of us can think of many others. For instance, as I presently am, I simply would be unable, not only to harm my daughters, but even to want to harm them. This is simply unthinkable for me. But I do not see why this deprives me of the praise I deserve, if I actually do, for caring about them and being a decently good father.

Things are very different, however, when someone acts in a morally wrong and objectionable way. Suppose that someone deliberately shows a vile and appalling attitude toward a weak and poor woman, insulting or even beating her. Now imagine that the aggressor utters the same words as Luther did: "I can do no other." In this case, unlike Luther's, we do not take naturally this avowal as a manifestation of the soundness and good functioning of the aggressor's rational faculties. Rather, we tend to assume that something is wrong with her, either in her rationality or in her moral sense, or both. And even if she adopts this attitude toward poor people as a result of her sincere belief in, say, a strong form of social Darwinism, we still do not tend to see her avowal, unlike Luther's, as a sign of her moral integrity, but rather as a symptom of her moral impairment or depravation. In this case, unlike Luther's, we assume that there *is* a different way of acting that she should have gone for; and we also assume that there is a different way of seeing things and of reasoning that she should have adopted. And, lacking evidence to the contrary, we assume that, since she *should* have had different beliefs and acted in a different way, she *could* have done so. *Given that she should, and could, have done otherwise but did not,* she is so far morally blameworthy. We assume, then, that APs are relevant, in fact required, for moral blameworthiness *as unfulfilled moral obligations that the agent was able to fulfill*. This is why APs are not required in cases of morally praiseworthy acts: Here, there are no unfulfilled duties, and so no alternative ways of behaving, in which the agent satisfies them, are necessary for the agent to be blameless or praiseworthy. This is not so with blameworthy actions. So, Dennett was wrong in concluding, from the fact that APs are not required for an agent's MR in "Luther," praiseworthiness cases, that PAP is false in all cases.

We have contended that, in the context of MR ascriptions, when we hold that an agent *should* have acted in a different way than she did, we naturally assume that she *could* have acted in that way. This assumption is an application of an old moral principle, namely, that "ought" implies "can" (OIC). This principle seems to me fully reasonable; its plausibility stands out especially in its contrapositive form: If someone is unable to *A,* then she is not morally obliged to *A*. If someone cannot swim, she is not morally obliged to jump in the water in order to save a drowning person. The general moral duty to help someone in danger of dying gets suspended in this particular context and for this particular person precisely because she was unable to fulfill it. As I wrote some years ago, "Denying it [OIC] means to find it acceptable to burden people with moral duties that they cannot discharge, and to hold them morally responsible for actions that it is not in their power not to perform. Too much moral weight to carry, it seems" (Moya 2006, p. 30). This principle is under some pressure nowadays, and part of the reason some people have for rejecting it is precisely their rejection of PAP. But it is not clear at all that PAP is false, at least concerning blameworthiness.

Some people take moral dilemmas to be counterexamples to OIC; the reason is that, in them, supposedly, people ought to do something despite being unable to do it. But I think this is not right. Consider Sartre's example of the young Frenchman who faces the dilemma between joining the Résistance against the Nazis and taking care of his elderly and ill mother. This dilemma does not refute OIC, in my view. The agent has a moral duty to join the Résistance; but he can fulfill this duty by joining the Résistance, as OIC states. On the other hand, he has a moral duty to take care of his mother; but he can fulfill this duty by taking care of his mother, again as OIC says. But I think he has *not* the moral duty to do both, that is, to join the Résistance *and* to take care of his mother simultaneously, precisely because doing that is impossible for him, as looks pretty correct and OIC implies. In general, "*S* ought to *A*" and "*S* ought to *B*" do not logically imply "*S* ought to both *A and B*". Rather than seeing the falsity of PAP as a reason for rejecting OIC, my advice is to see the truth of OIC as a reason for accepting PAP as well.

Now, in the context of our proposal, we can explain a fact that has intrigued me for some years, namely that PAP can be derived from OIC in the case of moral blameworthiness, but not of moral praiseworthiness. In the former case, the derivation can proceed thus:

> A person is morally blameworthy for A-ing only if she ought not to have A-ed (Premise *P*)
> A person ought not to have A-ed only if she was able not to *A* (OIC)
> Therefore,
> A person is morally blameworthy for A-ing only if she was able not to *A* (PAP, applied to blameworthiness, by transitivity of the implication)

The first premise asserts, plausibly enough, that blameworthiness implies a violation of a moral duty: The agent did something she should not have done (or omitted to do something she should have done). The second premise is OIC.

Let us try now the derivation of PAP from OIC in the case of praiseworthiness:

> A person is morally praiseworthy for A-ing only if she ought to have A-ed (Premise *P'*)

A person ought to have A-ed only if she was able to *A* (OIC)
Therefore,
A person is morally praiseworthy for A-ing only if she was able to *A*.

The conclusion in this case may be interesting: It suggests that A-ing by mere luck, even if A-ing is morally right, does not make someone praiseworthy. Interesting as this may be, it is not what PAP says, however.

These logical relations are illuminating of internal conceptual connections in the field of MR. PAP connects blameworthiness for A-ing (not A-ing) with ability not to *A* (to *A*); the relation may not be obvious, and it actually has been strongly contested; OIC connects moral obligation (not) to *A* with ability (not) to *A*; and Premise *P* provides the missing link by connecting blameworthiness for (not) A-ing with moral obligation not to *A* (to *A*). OIC and Premise *P* provide the rationale for PAP, by justifying the connection it establishes between avoidability and blameworthiness. And they also show why PAP cannot be derived from OIC for praiseworthy actions: Premise *P* is not true when we substitute "praiseworthy" for "blameworthy"; it is not true that an agent is praiseworthy for *A* only if she ought not to have A-ed; but this transit from *A* to not-*A,* which Premise *P* makes, is essential for PAP to be derived, since it also contains this relation. Praiseworthy actions, however, confirm OIC: Since the agent has done what she should, she clearly was able to do it.

2.3 Frankfurt Cases

Let us now attend to Frankfurt cases. They are probably the most important reason why many thinkers have been led to think that PAP is false (and to reject OIC as well). A typical Frankfurt case might be the following.

(ANN) Ann, a student, hates her colleague John and decides to lie to him about the date of an oncoming examination. As a result, John misses the exam, with some nasty consequences for him. Unbeknownst to Ann, Black, a clever and nefarious neurosurgeon who also wants Ann to lie to John, has implanted in her brain a device that allows him to monitor Ann's deliberation. If, on this basis, it is clear to Black that Ann is going to decide *not* to lie, he will press a button that will ensure that Ann will decide to lie. However, Black prefers not to intervene unless it is necessary. And it is not, for Ann, deliberating fully on her own and for her own reasons, decides to lie to John and Black never presses his button.

This example raises a strong intuition to the effect that Ann is fully morally blameworthy for lying to John even if, owing to the lurking presence of Black, she could not have avoided lying. If so, then PAP is false. A way of resisting this conclusion is to hold that some APs remain. Ann could, for example, have shown the sign that would have alerted Black of an oncoming decision to tell the truth instead of lying. Or she could have postponed her lie. However, not any AP goes in order to sustain PAP against Frankfurt cases. To use Fischer's terms, APs have to be *robust* in the sense of being explanatorily relevant to an agent's MR. Exactly what robustness or explanatory relevance includes is a matter of some controversy, but two aspects

seem to be essential to it.[2] First, in order to be robust, an alternative has to be under the agent's control. Suppose that Ann might have accidentally fainted before lying to John. It looks clearly wrong to try to save PAP by adverting to this alternative, even if it is present. Moreover, a robust alternative possibility has to be morally significant, in the sense that its presence is relevant to the question of whether the agent is morally responsible. So, for instance, though it was within Ann's control to lie to John in a lower or higher tone of voice than she actually used, this alternative is not relevant to her blameworthiness for lying. Some thinkers, such as Pereboom or Otsuka, insist on a third and very strong requirement, namely that robust APs have to be such that, had the agent gone for them, she would have been fully exempted from blameworthiness. Though quite demanding, the requirement has some plausibility.[3] In this chapter, I will accept this third requirement, at least for the sake of the argument.

Now, in the context of our proposal, when we explain an agent's blameworthiness by appealing to things she could have done but did not, we are presupposing that these are things that she *should* have done but did not. In our view, then, robust, morally significant APs are *unfulfilled moral duties*. Irrelevant APs, mere "flickers," to use Fischer's term again, are so because they are not things that the agent should have done to be exempted from blame. Our example of Ann's lying in a lower or higher tone of voice is a case at hand. Our account can give, then, a principled justification of the requirement of robustness and moral significance for those APs that, in Frankfurt cases, can be adverted to in order to defend PAP.

The intuitions raised by Frankfurt cases may be explained thus: If a rational and reasons-responsive agent deliberated, decided and acted fully on her own and for her own reasons, without being coerced or compelled, she is morally responsible for what she did; it does not matter for her MR whether, unbeknownst to her, there are factors in the situation that, without interfering at all in the process of decision and action, make it impossible for her to decide and act otherwise. So, why should we accept that someone is morally responsible for something she did only if she could have done otherwise? The supposedly obvious connection between MR and APs begins to appear as rather arbitrary and groundless.

Things look different, however, if, following our proposal, we introduce the concept of moral obligation. Instead of asking, in a situation where MR is at issue, whether the agent could have done otherwise, let us start by asking whether the agent *should* have done otherwise. This question has an immediate and direct impact on the agent's MR. If the answer is affirmative, so that the agent did not act as she should, then there is prima facie reason to think that she is blameworthy for what she did. Now, whereas saying, "She is morally blameworthy for what she did

[2] A good and useful characterization can be found in Capes 2010, esp., pp. 69–70.

[3] Pereboom may be right that the intuition behind the requirement of APs for MR is the "off the hook" intuition, the assumption that, in order to be blameworthy for some action, one must be able to act in a way that makes one fully blameless. This is essentially Otsuka's Principle of the Avoidability of Blame (cf. Otsuka 1998), a principle that he thinks respects the spirit of PAP and is not refuted by Frankfurt cases, even if PAP itself might be.

because she could have done otherwise" may be in need of further justification, saying, "She is morally blameworthy for what she did because she should have done otherwise" is not. In fact, it could be used as part of a justification of the former statement, for, if we hold that someone should have done other than she did, we are assuming, applying OIC, that she was able to do it.

When it is praiseworthiness that concerns us, things change significantly. Saying, "She is morally praiseworthy for what she did because she could have done otherwise" is not simply in need of further justification; it looks rather dubious, if not false, in view of "Luther" cases. The corresponding "should" statement would be, "She is morally praiseworthy for what she did because she did what she should"; and clearly, if we say this, we are *not* assuming that she could have done otherwise (even if she actually could). So, unlike blameworthiness, here we cannot use the "should" statement to justify the "could" statement.

How does all this affect Frankfurt cases? Concerning them, a primary question would be, "Did the agent do what she should, morally speaking?" If the answer is positive, then, ceteris paribus, the agent is morally praiseworthy, and the question, "Could she have done otherwise?" does not change this positive assessment.

Suppose, however, that the answer to that primary question is negative; then there seem to be unfulfilled moral duties, things that the agent should have done but did not. In such a case, the question whether she could (was able to) have done those things becomes crucial. If she could, then she had within her reach robust, even exempting, APs, unfulfilled moral duties such that, had she fulfilled them, her blame would have decreased or even disappeared. Suppose, however, that the agent could not have done the things she should have done. In this case, *given OIC,* there were no unfulfilled moral duties after all, for on this occasion they were suspended for the agent by virtue of her impossibility to satisfy them. But then she is not blameworthy, for there was nothing that she should have done but did not.

In either case, then, PAP, applied to blameworthiness, is confirmed. In the first case, in which the agent could have fulfilled her duties, she is blameworthy, but she had robust, even exempting, APs. In the second case, she had no robust APs, but she was not blameworthy, either.

Frankfurt theorists will object to the preceding considerations on the following lines. Assuming the truth of OIC is sort of question begging against them, for they think that PAP is false, so that there can be blameworthiness without robust APs; and if OIC, together with an additional premise, implies PAP, OIC may be false as well. So, the line of argument that has taken us to deny the agent's blameworthiness is defective. According to them, some Frankfurt cases feature agents who could not act so as to get exempted from blame but are blameworthy for what they did anyway.

At this point, a response might be to insist on the independent plausibility of OIC and on the unpalatable consequences of denying it. I think these considerations are fully correct. But it is unlikely that this line of argument can convince Frankfurt theorists. They will point to the intuitions raised by the relevant Frankfurt cases in order to reject both PAP and OIC and to show that denying the latter principle has not the unacceptable consequences one could initially expect. We would be caught in a dialectical stalemate.

2.4 Doing Everything One Can (DEC) and Frankfurt Cases

My suggestion at this point is to resort to a new principle, drawn from common sense wisdom, which has some connections with OIC but is slightly different from it. Moreover, it is more intuitively correct than OIC. It is different because, whereas OIC relates two notions, namely moral obligation and ability or power, the new principle connects three notions, namely moral obligation, ability or power and blameworthiness. That it is more intuitive than OIC is something to be left to the reader's opinion. Let me call this principle "DEC" (from "doing everything one can"):

> (DEC) If someone does everything she can reasonably do in order to fulfil her moral duties, she is not obliged to do more, and so is not morally blameworthy for not doing more.[4]

DEC sounds eminently correct and in full agreement with our sense of justice. It seems clearly unfair to demand from an agent to do more when she has done everything he can reasonably do in order to comply with morality. Let us now apply this principle to some examples, including Frankfurt cases.

An important feature of Frankfurt cases is that, in them, the agent believes that she can do otherwise than she in fact does, though, owing to circumstances fully unknown to her, this belief turns out to be false. A common situation with this feature might be the following:

> (JOHN) In his way home from work, John meets Robert, a friend of his; Robert tells John that he urgently needs to take a taxi to get to an important meeting in time and asks him for some money, since he has realized he does not have enough cash with him; of course, he goes on, he will return the money as soon as possible. John thinks he can comply with Robert's request quite easily, but he is in rather bad temper and not in the mood to be kind; and so he responds to Robert that he is sorry but, unfortunately, he has no money with him; he forgot to take his wallet this morning while leaving home. Unknown to John, however, he actually has no money and no wallet, for a pickpocket stole it half an hour ago while John was travelling in the underground train.

Is John blameworthy for not helping Robert with the money he needed? Well, not exactly for that. If someone censured him for not helping his friend, he might reply that he is not to blame for that, since he could not have helped him. But he is not fully blameless, either. Following DEC, he did not do everything he could reasonably have done for helping his friend. He should have formed the intention to help and tried to do so by reaching for his wallet, if only to discover that it was not in his pocket. Had he done so, and given that this is everything he could reasonably have

[4] Note that it does not follow from DEC that anyone who fails to do everything she reasonably can do to fulfill her moral duties is ipso facto morally obliged to do more and morally blameworthy for not doing it, though such a failure provides a prima facie reason for thinking that she is. I believe that this is how it should be, for otherwise DEC would often put on our shoulders too heavy moral burdens, given that in many cases we can do more, even if we do much. The qualification "reasonably" is important, as well as it is attending to the details of particular cases.

done to help his friend, he would have been entirely blameless.[5] In this particular context, his attempt to help by reaching for his wallet was a fully robust, exempting alternative.

But let us change a bit the situation. Suppose that no robbery has occurred and John has in fact his wallet in his pocket. In these circumstances, reaching for his wallet would not have been enough to get rid of blame, for then there would have been something more John could have done to fulfill his duty; in the new situation, only lending the money to Robert would have made John entirely blameless. So, an act of a certain kind (John's forming the sincere intention to lend the money, manifested in his reaching for his wallet) can be an exempting alternative in a certain situation, provided that this is everything the agent could do to fulfill his obligation, but not in another, if there is something more he could have done.

A general lesson to draw might be this: In situations in which MR is at issue, the more an agent can do to fulfill her moral duties, the more she must do to get rid of blame; conversely, the less she can do, the less she must do to be blameless. So, which kinds of actions are fully robust, exempting APs is a contextual matter, and depends importantly on the effective extension, in the particular situation an agent is, of her power or ability to fulfill her moral obligations.

Now, Frankfurt cases are precisely examples of situations where an agent's powers are in fact, unbeknownst to her, severely restricted. In them, and unlike John's example, the agent cannot even form an intention or make a decision to act in a certain (morally right) way, as happens in Ann's example. Nevertheless, Frankfurt theorists insist that she is blameworthy for what she intends or decides (and does) in spite of having no robust alternative intention or decision, not to mention action, within her reach; she is blameworthy even if there is no alternative way of acting that would render her blameless. If this is correct, then PAP is false not only with respect to praise, but to blameworthiness as well.

In the light of our preceding line of argument, however, this thesis can be challenged. Let me proceed.

Ann's example is a classical Frankfurt case. However, as a result of the discussion, cases of this sort are currently taken to be defective, though we shall not get into the reasons for this view.[6] This has led Frankfurt theorists to construct new and more complex cases, which supposedly do not show the unwanted features. The following is a Frankfurt case of this new brand, which combines features of some recent Frankfurt cases (such as Pereboom's *Tax Evasion* (Pereboom 2001, 2009) or Widerker's *Brain Malfunction-W* (Widerker 2006)):

> (MEAN): Frank is a successful businessman; he occupies a relevant position in an important company and has a pretty good financial situation. His brother Neal, instead, is in a rather precarious economic condition. One day, Neal visits Frank to tell him he has had to incur some debts, not very high indeed, but which can cause him real trouble if he does

[5] Reasonably: John could, for example, commit a robbery to get the money Robert needed; but this would not have been a reasonable thing for him to do.

[6] The main problem is the so-called dilemma defence of PAP, which has been put forward by such thinkers as Kane (1985), Widerker (1995), and Ginet (1996).

> not repay them within the next few days. So, Neal asks Frank to lend him money to cancel the debts. Frank is a rather mean and egoist person, though not pathologically so. Frank's psychology is such that, in a circumstance like this, he only could decide to lend the money to Neal if he took seriously moral reasons to do so; but taking seriously these reasons is not sufficient for him to make this decision; even if he took these reasons seriously, he still might refuse to help Neal for reasons of self-interest, related to his mean and egoist character. Unbeknownst to Frank, however, he could not even make the decision to help Neal, for his present health condition (which he is ignorant of) is such that, if he were to take seriously moral reasons to help Neal, and with them the perspective to do actually so, he would suffer a grave heart attack, with the result that he would lose consciousness and be taken to the hospital, which would make it impossible for him to lend Neal in time the money he badly needs. However, no heart attack takes place, for Frank does not take seriously moral reasons to help his poor brother and refuses to lend him the money by inventing an excuse about his own very precarious financial situation.

According to Frankfurt theorists, Frank is blameworthy for refusing to help Neal, though he could not have decided to help him. They concede that Frank had within his power taking seriously moral reasons to help Neal; this is an alternative he had, something he did not but was able to do. But this alternative, they contend, is not robust, in that it is not morally significant. It is true that, had he gone for it, he would have been exempted from blame for not helping Neal, because, by suffering a heart attack, he would have been unable to help him. But this would have happened by pure luck; Frank could not reasonably believe or foresee that, simply by taking seriously moral reasons, he would be blameless; remember that, barring Frank's present (and passing) health condition, which had no actual role in his final decision, taking such reasons seriously was fully compatible with the decision to let Neal down anyway.

With examples such as this, Frankfurt theorists seem to occupy a very solid position indeed. And we must acknowledge that MEAN (and the original cases that inspire it) is a strong example against the necessity of robust APs for blameworthiness. If thinking of moral reasons were a robust alternative, it would be very easy to get off the moral hook each time we did a morally objectionable thing, such as refusing to help a friend or a relative, as Frank did in MEAN. We might excuse ourselves in the following way: "Look, I have decided not to lend you the money you need, but I am not blameworthy because I have taken seriously moral reasons for lending you the money." But this is clearly ludicrous. Nobody could think of this as a valid excuse.

This is undoubtedly correct, but we can give a strong reply to it. Remember our initial example of John and Robert. We said that which ways of acting are robust, exempting APs is a contextual matter, and depends especially on what an agent can effectively do in particular circumstances. This is why if, unknown to John, he had been stolen and had no money, a sincere, though unsuccessful, intention to lend Robert the cash he needed would have exempted John from blame. But if no robbery occurred and John actually had money, the mere intention would not have been enough to exempt him from blame; only lending the money would have had this effect. By the same reasoning, if we can actually help a friend in trouble, we cannot get rid of blame for not helping him merely by taking seriously reasons for helping him. We can then explain why, in general, taking seriously moral reasons

for acting in a certain morally right way is not robust enough to get rid of blame for not deciding and acting that way. But this does not mean that taking moral reasons seriously is not a moral duty in a situation like the ones we are considering. We are rightly expected to think of our friends' wellbeing and put ourselves in their shoes, as a step toward intending and doing the right thing for these right moral reasons. Failure to attend to considerations of this kind is a reason for blaming someone as selfish or inconsiderate.[7]

Coming back to MEAN, remember that, unknown to Frank, he could not actually help Neal; he could not even intend to help him; the only moral duty he was actually able to fulfill was precisely to think of Neal's troubles, put himself into his shoes and consider seriously reasons for helping him. So, by parity of reasoning, in the same way as, following DEC, John would have been exempted from blame by a sincere intention to help Robert if, due to the robbery, he was actually unable to help him, Frank would also have been exempted from the blame he now has for refusing to help Neal if he had taken seriously moral reasons for helping him, given that this was the only thing he could in fact have done, in the circumstances, to fulfill his moral duty to help his brother Neal. Frank did not satisfy the antecedent of the DEC principle; he did not do everything he could reasonably have done in order to fulfill his moral duties; he had an unfulfilled moral obligation which he could have fulfilled and did not; there is, then, reason to think that he is morally blameworthy for his refusal to help Neal. In the circumstances of the case, the tiny, though morally significant, mental act of taking moral reasons seriously was a robust, in fact exempting, alternative possibility that Frank had within his power. Therefore, MEAN and similar, highly sophisticated Frankfurt cases do not show that, in the absence of any robust, morally significant, exempting alternative, an agent can be blameworthy for what she did. Frank was blameworthy but he had such an alternative. MEAN has not shown that blameworthiness does not require APs and PAP, concerning blameworthiness, remains safe.

References

Capes JA (2010) The W-defense. Philos Stud 150:61–77

Dennett D (1984) Elbow room: the varieties of free will worth wanting. Clarendon Press, Oxford

Fischer J (1994) The metaphysics of free will. Blackwell, Oxford

Frankfurt H (1969) Alternate possibilities and moral responsibility. J Philos 66:829–839. (Reprinted in (and quoted from) Frankfurt H (1988) The Importance of What We Care About. Cambridge University Press, Cambridge)

Ginet C (1996) In defence of the principle of alternative possibilities: why I don't find Frankfurt's argument convincing. Philos Perspect 10:403–417

Kane R (1985) Free will and values. SUNY, Albany

[7] I am aware of Watson's (2004) important distinction between two aspects of responsibility; but getting thoroughly into this would take us too far away from our main line of argument. I assume that someone who is inconsiderate and refuses to help a friend in trouble is blameworthy in both aspects that Watson distinguishes.

Moya C (2006) Moral responsibility. The ways of scepticism. Routledge, Abingdon
Naylor M (1984) Frankfurt on the principle of alternate possibilities. Philos Stud 46:249–258
Nelkin D (2008) Responsibility and rational abilities: defending an asymmetrical view. Pac Philos Q 89:497–515
Otsuka M (1998) Incompatibilism and the avoidability of blame. Ethics 108:685–701
Pereboom D (2001) Living without free will. Cambridge University Press, Cambridge
Pereboom D (2009) Further thoughts about a Frankfurt-style argument. Philos Explor 12:109–118
Watson G (2004) Two faces of responsibility. In: Watson G (ed) Agency and answerability: selected essays. Clarendon Press, Oxford, pp 260–288
Widerker D (1995) Libertarianism and Frankfurt's attack on the principle of alternative possibilities. Philos Rev 104:247–261
Widerker D (2006) Libertarianism and the philosophical significance of Frankfurt scenarios. J Philos 103:169–187
Wolf S (1990) Freedom within reason. Oxford University Press, Oxford
Wyma K (1997) Moral responsibility and leeway for action. Am Philos Q 34:57–70

Chapter 3
The Normativity of Evaluative Concepts

Christine Tappolet

Abstract It is generally accepted that there are two kinds of normative concepts: evaluative concepts, such as *good,* and deontic concepts, such as *ought*. The question that is raised by this distinction is how it is possible to claim that evaluative concepts are normative. Given that deontic concepts appear to be at the heart of normativity, the bigger the gap between evaluative and deontic concepts, the lesser it appears plausible to say that evaluative concepts are normative. After having presented the main differences between evaluative and deontic concepts, and shown that there is more than a superficial difference between the two kinds, the chapter turns to the question of the normativity of evaluative concepts. It will become clear that, even if these concepts have different functions, there are a great many ties between evaluative concepts, on one hand, and the concepts of ought and of reason, on the other.

Keywords Normativity · Evaluative concepts · Deontic concepts · Good · Ought

One can say without exaggeration that normativity has become one of the central themes in contemporary philosophy, if not *the* central theme. But what is normativity, exactly? Paradoxically, this is a somewhat neglected question. As Kevin Mulligan points out, a great number of ordinary concepts that are considered to be part of the same family, which we have acquired the habit of qualifying as 'normative':

> We may say of a particular action performed by Sam that it is *elegant* or *evil*, that he *ought not* to be doing what he is doing, that it is the *right* thing to do, that he is *obliged* to do it, that it is his *duty*, that he has a *right* to act as he does, or that it is *virtuous*. The different properties we ascribe in this way belong to one very large family which, for want of a better word, we may call *normative* properties. (Mulligan 2009, p. 402)

The concepts that count as normative can appear quite heterogenous. However, some groupings seem natural. Thus, it is generally admitted that we can divide these concepts into two large distinct groups: evaluative or axiological concepts (from the latin *valores* or the Greek *axos,* both meaning that which has worth), such as *good* and *bad,* and deontic concepts (from the Greek *deon,* meaning that which is binding), such as *obligatory* and *permissible* (von Wright 1963; Wiggins 1976; Heyd 1982;

C. Tappolet (✉)
Université de Montréal, Montréal, Canada
e-mail: christine.tappolet@umontreal.ca

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_3,

Thomson 1992, 2007, 2008; Mulligan 1989, 1998, 2009; Dancy 2000a, b; Smith 2005; Ogien and Tappolet 2009; Wedgwood 2009).

The distinction between evaluative and deontic consists in a generalisation of the traditional opposition between the good and the right. The question of the relation between the evaluative and the deontic has been the object of numerous debates. Thus, one of the central tasks for all who are interested in ethics, but also in epistemology, aesthetics or any normative domain, is to specify the relation between these two families of concepts. Is one of the kinds more fundamental, conceptually speaking, than the other? That would mean that to possess one of the two kinds, one would have to possess the other. If there were such an asymmetry, which of the two kinds of concepts would be the more fundamental? Would the possession of deontic concepts be necessary for the possession of evaluative concepts or would it be the other way round, with evaluative concepts as more fundamental? A third possibility is to deny that one of the two kinds of concepts is more fundamental than the other. The two kinds of concepts would be at the same level, conceptually speaking.[1]

It is worth noting that apart from the question of conceptual priority, other questions of priority arise.[2] One can raise the metaphysical question by asking not only about the priority of evaluative and deontic concepts, but also about the objects that seem to correspond to them, whether these are properties or not. More generally, this question involves the relation between evaluative facts and deontic facts, supposing that these two types of fact exist. Finally, the question of priority also suggests itself when talking of explanation. Are evaluative facts able to explain deontic facts or, vice versa, can deontic facts explain evaluative facts? Evidently, one cannot rule out the possibility that nothing explains the facts in question, or that there are another kind of facts, such as, perhaps, natural nonnormative facts, which can explain both evaluative and deontic facts.

The question that interests me here lies at the conceptual level. It is the question of whether it is possible to claim that evaluative concepts are normative. More precisely, if one maintains that evaluative and deontic concepts belong to two distinct conceptual families, how is it possible to consider evaluative concepts well and truly normative? In fact, it is plausible to claim that deontic concepts, more particularly the concept of ought, are at the heart of normativity. Therefore, the wider the distance between evaluative and deontic concepts, the lesser it will seem true that evaluative concepts are a kind of normative concepts. In a more general manner, the question of the unity of the normative domain is at play here. Indeed, the division into two distinct groups raises the question of how it can be true that two kinds of concepts belong to one and the same family.

The thesis according to which there is a real, rather than superficial, difference between evaluative and deontic concepts have been the object of criticism. For reasons which seem principally strategic, a uniform treatment of the normative domain has seemed particularly seductive for those who subscribe to prescriptivism, the doctrine according to which moral judgements are assimilated to imperatives.

[1] This is what Wedgwood (2009) maintains.

[2] For the distinction between these three kinds of normativity, see Väyrynen (2010).

Rudolf Carnap formulated the most striking rejection of the distinction between the evaluative and the deontic. According to Carnap, the difference between an evaluative judgement, such as 'Killing is bad' and a norm or a rule, such as 'Don't kill', is merely one formulation. In fact, both statements have an imperative form and are neither true nor false. For Carnap, 'a value statement is nothing other than a command in a misleading grammatical form' (1935, p. 24). Carnap's conception is close to that of Richard Hare (1952). So, although Hare mentions a certain number of differences on the level of 'grammatical behaviour' between 'good' on the one hand and 'right' and 'ought' on the other, he maintains that there is enough similarity between 'good' and 'right' and 'ought' to consider all three evaluative (Hare 1952, pp. 152–153). It would be false to conclude that Hare thinks that evaluative concepts have priority. In fact, according to the classification that he proposes at the beginning of his book, imperatives as evaluative judgements form part of a larger class of 'Prescriptive Language' (Hare 1952, p. 3, 153). As becomes clear at the end of his book, Hare in fact maintains that statements containing 'right' and 'good' can be replaced by statements containing 'ought'. Statements containing 'ought' can, in turn, be formulated in the imperative form (Hare 1952, pp. 180–181).

My plan is the following. To measure the gap that separates evaluative concepts from deontic concepts, I will begin by presenting the principal differences between the two kinds of concepts (for a more complete presentation, see Ogien and Tappolet 2009, Chap. 2). Indeed, the question is whether there is more than a merely superficial difference between the two kinds of concepts and, if that is the case, what this difference consists in. After this, I will turn to the question of the normativity of evaluative concepts.

Before beginning, I should make a point about my methodology. The truth is that there is no agreement over which terms count as evaluative and which as deontic. For example, is the concept of reason, in the normative sense of the term, evaluative or deontic, supposing that it falls into one of the two categories?[3] Faced with this difficulty, the best strategy is to work principally with paradigmatic cases, such as *good* and *bad* for evaluative, and *ought* for deontic. It is only by grasping cases of this kind that it will be possible to deal with the difficult cases, such as the concept of reason.

3.1 The Gap Between the Evaluative and the Deontic

The first reason for distinguishing between evaluative and deontic concepts is that these concepts form two distinct and 'tightly knit' conceptual families. On the one hand we have the family organised around the concepts *good* and *bad,* but which also includes the concept *indifferent*. On the other hand, we have the family made up of *obligatory, permissible* and *forbidden,* which constitute the domain of deontic logic.[4]

[3] Smith (2005, p. 11), for example, counts reason as a deontic concept.

[4] A more complete list includes *gratuitous* and *optional* (McNamara 2006).

The members of each of these families are connected by direct inferential links. If something is good, it follows that it is not bad. In fact, the three most general evaluative concepts seem interdependent. What is good is what is neither indifferent nor bad; what is bad is what is neither indifferent nor good; and what is indifferent is what is neither good nor bad. These links seem to form part of what we learn when we acquire the concepts in question. The assertion 'If something is good, it is neither bad nor indifferent' is one of the truisms describing the dispositions to make inferences that characterise possession of these concepts.

By comparison, the relation between evaluative and deontic concepts seems slacker. It is possible to maintain that evaluative concepts can be analysed or elucidated with the help of deontic concepts or, vice versa, that deontic concepts can be analysed or elucidated with the help of evaluative concepts, or even that the two kinds of concepts can be analysed or elucidated by a third kind of concepts. For example, according to a suggestion tracing back to Franz Brentano (1889) and that has recently been the object of renewed interest, evaluative concepts can be analysed or elucidated with the help of the notion of appropriate (or fitting) attitudes, a notion which many authors consider deontic (among others, Smith 2005; Schroeder 2008; Bykvist 2009). Thus, it would be true that something were good if and only if it were appropriate to have an attitude of approbation towards it. Whether or not such a suggestion is deemed overall defensible, it seems clear that one cannot consider it a simple truism. One has only to think of the debates this kind of conception has inspired to convince oneself of this. The same point applies to the inverse suggestions, according to which deontic concepts can be analysed or elucidated with the help of evaluative concepts. Thus, the assertion, dating back to G.E. Moore (1903), according to which one must carry out an action if and only if this action is that which has the best consequences, or simply is the best of all possible actions, is perhaps true, but it is certainly not a truism.

A second consideration that allows us to differentiate evaluative and deontic concepts concerns the number of elements in each of the two families: The evaluative family is much bigger than the deontic family (Mulligan 1989, 1998, 2009). As many have pointed out, the concept *good* (and also *bad,* of course) can be used in a variety of ways (Ross 1930, p. 6; von Wright 1963; Thomson 1992, 2008; Wedgwood 2009). Something can be called good *simpliciter,* such as when one says that knowledge or pleasure is good. When one says this kind of things, one uses the term 'predicatively', as an authentic predicate, and not 'attributively' as a term modifying a predicate (Ross 1930, p. 65; Geach 1956, p. 33). And yet *good* can also be used attributively. One can also say that Sophie is a good philosopher, but a very mediocre cook. Furthermore, there are various locutions that involve the term *good* (von Wright 1963; Thomson 1992; Wedgwood 2009). Indeed, one can say that something or someone is good for something or for someone, or that something is good for an end, or even for doing something. In each case, it seems that the thing or person in question is good in a way, to use Judith Thomson's expression (1992).

The family of evaluative concepts also includes more specific concepts, such as *admirable, desirable, fair, generous, honest, benevolent* and *courageous,* to name

only a few of the terms used to express approbation.[5] It is fitting to stress that the attribution of these terms, in their ordinary usage, implies the attribution of the more general evaluative concepts, *good, bad* and *indifferent*. So, the question of whether what is admirable is also good or, more exactly, good in a way, does not arise. If an action is admirable, it is necessarily good from this point of view.[6]

By comparison, the family of deontic terms is much poorer. There does not seem to be a specific way of being obligatory, permissible or forbidden. It is true that one can distinguish between different kinds of obligations: moral obligations, legal obligations and prudential obligations seem well and truly distinct. However, even if one allows that there are different ways of being obligatory, rather than the same idea of obligation applied to different domains—which is far from evident—one must recognise that the deontic family is still poorer than the evaluative family. Indeed, the latter also engages with different normative domains, so that one can distinguish what is good from a moral point of view from what is good from a legal or prudential point of view.

The thesis according to which the deontic family is poorer than the evaluative family has been recently criticised by Ralph Wedgwood (2009). He claims that the English terms *ought* and *should* are comparable with 'good' in the sense that they are multivocal. They are capable of expressing many different concepts in different contexts. Wedgwood distinguishes four kinds of *oughts*. The first, which he calls the '"ought"' of general desirability', is what one uses when one says 'Milton ought to be alive', or 'there ought to be world peace'. It is what ought to be, as opposed to what an agent should do, that is at play here.[7] The second is the *ought to do*, which Wedgwood calls the 'practical "ought"'. This kind of *ought* is indexed to a particular agent at a particular time and involves the actions that the agent in question is capable of accomplishing. The third kind of *ought,* qualified as relative to an end and that one could call 'instrumental "ought"', is illustrated in the statement 'He ought to use a Phillips screwdriver to open that safe'. Finally, the fourth kind of *ought,* which is qualified as 'conditional "ought"' , concerns what one ought to do when one does not do what one ought really to do. This usage is illustrated in 'If you don't stop shooting up heroin, you ought at least to use clean needles', where it is understood that one ought to stop shooting up heroin.

Should we conclude that the deontic family is as large as the evaluative family? I think not. An initial question arises concerning the notion of ought to be. Indeed, the fact that Wedgwood is tempted to talk of the *ought of general desirability* is

[5] Many of these concepts are called 'thick' evaluative concepts, in contrast to 'thin' evaluative concepts (Williams 1985). Thick concepts are characterised by the fact that they include a purely descriptive element. For example, the attribution of the term 'courageous', implies an attribution of the capacity to stand up to danger, or more generally to difficulties. On the basis of this distinction, one can say that, contrary to deontic concepts, evaluative concepts can be thick (Mulligan 1998, pp. 164–165).

[6] This is what Wallace (2010) fails to recognise.

[7] As Wedgwood recalls, Sidgwick ironically talks of the 'political ought' to designate this kind of ought. Mark Schroeder (2011) qualifies this notion of evaluative ought and distinguishes it from what he calls the deliberative ought.

evidence of this; one can question whether he is really discussing a deontic notion here.[8] A second question involves the relation between different usages of the term 'ought'. Could we not reduce the instrumental and conditional *oughts* to practical *oughts*? Nevertheless, let us suppose that there really do exist four kinds of distinct *oughts*. It would still be true that the family of evaluative concepts is much more numerous. Four usages are very little in comparison with the multitude of usages of *good* and *bad*. Furthermore, there is an important difference regarding the structure of the two conceptual families. As we have seen, the evaluative family includes general and specific terms, which does not seem true of the deontic family. None of the four *oughts* listed by Wedgwood is more general than the others.

Another point we should note in this context is that evaluative concepts, particularly some of the more specific evaluative concepts, are closely tied to affective reactions (Mulligan 1989, 1998, p. 166). Concepts such as *admirable* or *contemptible,* which correspond to words that are lexically tied to affective terms, are the first to spring to mind; but it also seems plausible to think that more general evaluative concepts, such as *good* and *bad,* are tied to specific affective reactions—approbation and disapprobation—or even an ensemble of affective reactions—positive reactions and negative reactions. In contrast, the relation between deontic concepts and affective reactions seems much less tight. There is no lexical relation between 'obligatory', 'permissible' and 'forbidden', on the one hand, and terms that reflect affective reactions, on the other. More generally, no specific emotion seems to exist that corresponds to the obligatory, nor to the permitted, nor to the forbidden.

A third consideration weighing in favour of the existence of a real distinction between evaluative and deontic concepts is that evaluative concepts, but apparently not deontic concepts, can take comparative and superlative forms (Hare 1952, p. 152; Mulligan 1998; Wedgwood 2009). In other words, values, but not *oughts,* admit of degrees. One can say of someone that she is more or less admirable, or that her action is more or less courageous. And one can also say that a novel is better than another. Ordinary deontic terms, on the other hand, do not seem to allow comparative and superlative forms. As Hume noted, one does not say that something is more or less obligatory, or else that an action is more forbidden than another (1739–1740, pp. III, vi: 530–531). A plausible explanation of the absolute nature of deontic concepts is that these concepts are applied primarily to things that do not admit of degrees, that is to say, actions. Actions can be characterised by all kinds of properties that admit of degrees—one can sing more or less loudly or more or less out of tune—but one has to either act or not act—in principle, there is no way of *more or less* singing: one either sings, or one does not (Ogien and Tappolet 2009, pp. 64–65). This is a particularly important point in the context of deliberation or decision. When you try to work out what to do, you need to know whether a particular action ought or ought not to be performed. The conclusion that an action is one that one ought to perform to certain degree—it is *a bit* obligatory to perform it—is not what is sought.

[8] As I have already remarked, Mark Schroeder talks of 'evaluative ought'.

It could be objected that we implicitly allow for deontic comparisons when we conceptualise moral dilemmas. Suppose that an agent has a choice between killing or lying. We will certainly conclude that this agent ought to lie rather than kill. Thus, one can ask whether this is not the same as saying that one ought to lie more than one ought to kill, or that killing is more forbidden than lying. It seems in any case that the prohibition on killing has priority over the prohibition on lying (Hansson 2001). Furthermore, ordinary language seems to allow for deontic nuance. We distinguish between what *must* be done and what *should* be done, for example (Hansson 2001, pp. 131–132; Thomson 2008, p. 124, 229–230). Should we therefore think that, despite appearances, deontic concepts do admit of degrees? No, because we should recognise that the existence of a relation of priority between different *oughts,* something which is hard to deny when there is no question of doing all the considered actions, does not imply the existence of a relation of degree (Mulligan 1998, p. 164, for the idea that recognising that one promise binds us more than another does not imply deontic degrees).

A fourth consideration that can be put forward to support the claim that there is an important difference between evaluative and deontic concepts concerns the logical form of evaluative and deontic statements (Mulligan 1989, 1998). At first glance, the simplest evaluative judgements, such as 'this is good' have a subject-predicate form, $F(x)$, where the evaluative terms stand for predicates. Deontic concepts, on the other hand, are standardly taken to be propositional operators, which means that deontic judgements are taken to have the form $O(p)$ (where 'O' stands for obligatory).

However, things are not so straightforward. Firstly, evaluative terms can take the form of propositional operators, such as when we say that it is good, or desirable, that it rains. Secondly, we cannot rule out the possibility that the apparent structure of evaluative judgements is misleading. Their logical form could, for example, contain a tacit reference to a speaker or a social group. Moreover, deontic judgements can also take a variety of forms, such as when one says that doing this or that is forbidden, or that someone should do this or that. Finally, there is a reason to think that the hypothesis that deontic statements involve propositional operators is problematic. As Peter Geach (1991, p. 35) has argued, the hypothesis does not acknowledge that obligations concern agents and not just states of affairs. Geach maintains that deontic terms are operators taking verbs to make verbs. Thus, when we say that Sophie ought to sing, what we say is that *ought to sing* is true of Sophie.[9]

However, there nonetheless appears to be two important facts that distinguish evaluative from deontic judgements. The first is that some evaluative judgements resist transformation into judgements involving a deontic propositional operator. This is true not only of specific judgements like 'This is an admirable knife' or 'She is courageous', but also of sentences with more general evaluative terms, such as 'This soup is good for him.' In contrast, it appears that all deontic judgements can

[9] Mark Schroeder (2011) defends a similar thesis. Schroeder, who, contrary to Geach, argues that there are two kinds of *oughts,* deliberative *oughts,* relative to what is to do, and evaluative *oughts,* relative to what ought to be, claims that deliberative *oughts* reflect a relation between an agent and an action. In our example, the term 'ought' would reflect a relation between Sophie and the action of singing.

be transformed either into judgements involving a propositional operator or into judgements involving an operator modifying a verb. The other difference is that evaluative terms describing actions, but not deontic terms, can be transformed into adverbs that describe how an action is performed.[10] Suppose that Sally's action was both courageous and morally obligatory or required. We can say that Sally acted courageously, thus describing how she acted; but, even though in a sense she might be said to have acted obligatorily, we do not describe *how* she acted if we say this. There thus appears to be a category mistake involved in the sentence 'Sophie acted courageously, energetically and obligatorily.' Acting in the way you ought does not appear to be a way of acting. What these points suggest is that, in contrast to deontic concepts, evaluative concepts correspond to properties characterising things and people.

The next consideration that weighs in favour of a distinction between evaluative and deontic concepts concerns their respective domains of application. As David Heyd (1982, pp. 171–172) claims, it is clear that all sorts of things, ranging from persons and their actions to objects and states of affairs, can be the object of evaluation. In contrast, deontic concepts typically concern agents and their actions. It might thus be thought that deontic concepts only apply to what is subject to the will.[11] As expressed in the principle '*ought* implies *can'*, it is only as far as an agent is able to perform an action that she can be subjected to an obligation to perform that action. In fact, the domain of deontic concepts is broader, for it includes things such as beliefs, intentions, choices, emotions and character traits, etc. One can certainly say that a person should or should not believe something, have a certain intention, make such and such a choice, feel a certain emotion, possess such and such character trait, etc. And yet, it is often claimed that these things are not subject to the control of the will. Nonetheless, in as far as it is possible for an agent to have an indirect influence on her beliefs, intentions, etc. one can say that deontic concepts are concerned with things that have to be at least indirectly subject to the will.[12]

This claim poses a problem when it comes to judgements about what ought and ought not to be. These appear to be *bona fide* deontic judgments, but they are far from being concerned with things that are subject to the will, directly or indirectly. One could suggest that what ought to be should at least be possible (Wedgwood 2009, for this suggestion). But that is not certain. Indeed, if one accepts it is the best of all worlds that ought to be, and one also accepts that the world would be better if 2 + 2 made 5—this would allow us to feed more people, after all—one would have to conclude that an impossible world, even a logically impossible world, ought to be.

Nevertheless, it remains true that, compared to deontic concepts, evaluative concepts have a much broader diet. Evaluative concepts are omnivorous, while deontic

[10] This is the test proposed by Ogien and Tappolet (2009, p. 56).

[11] This would explain why it seems that deontic judgements imply the possibility of holding someone responsible (Smith 2005).

[12] Cuneo distinguishes between what he calls 'responsibility norms' and 'propriety norms', which apply not only to voluntary actions, but also to things that are beyond our direct voluntary control (2007, p. 82).

concepts are used either for that which is directly or indirectly subject to the will, or for states of affairs.

A final consideration in favour of the distinction between evaluative and deontic concepts concerns the possibility of dilemmas.[13] In contrast to evaluative judgements, deontic judgements seem to give rise to authentic dilemmas. As we know, our obligations can conflict, in the sense that we ought to perform two actions that are incompatible. If twins are drowning, it seems that one ought to save one as much as the other, even if it is impossible to do both because the twins are too far from one another. What we have in this kind of dilemma can be described in the following manner (to simplify, I will use propositional deontic operators) (Williams 1965; Tappolet 2004):

1. $O(p)$
2. $O(q)$
3. Impossible (p and q)

Of course, there are also value conflicts. It can be just as desirable to spend one's holidays by the sea as to spend them in the mountains, but it is unfortunately impossible to spend them in two places at the same time. Here is how we can formalise these conflicts (V is for value):

1. $V(p)$
2. $V(q)$
3. Impossible (p and q)

The difference between the two kinds of conflicts is that the first threatens to produce a contradiction, while the second does not. Indeed, the two principles that allow us to derive a contradiction—the principle that *ought* implies *can* and the principle of agglomeration—seem plausible in the case of obligations, but not in the case of values. As we have seen, it is plausible that *ought* implies *can*. It is only in as far as an agent is capable of fulfilling a requirement that this requirement can apply to him. The evaluative equivalence of the principle is clearly false: Something can be desirable or good while being impossible. Indeed, many things are. Moreover, as Bernard Williams (1965) suggested, the principle of agglomeration, although it seems plausible in the case of obligations, has no plausibility in the case of values. Indeed, it seems legitimate to say that someone who ought to keep her promise to Pierre, but ought also keep her promise to Paul, ought to keep her two promises. Contrary to this, it is easy to imagine that, even if doing something is good or desirable—marrying Pierre, for example—and that doing something else is also desirable—marrying Paul, for example—doing both things is not at all desirable: Marrying both Paul and Pierre might turn out to be a nightmare (supposing it were a legal possibility, of course). It is for this reason that some deny that it is possible that two obligations, or at least two obligations that are *all things considered,* can conflict. On the other hand, no one is tempted to deny that two incompatible things can be good, even good all things considered.

[13] This is a point that has recently been added to the list in Ogien and Tappolet (2009).

In summary, there are good reasons to think that there is more than a superficial difference between evaluative and deontic concepts. The two kinds of concepts each form a distinct conceptual family, linked by a cluster of truisms. The evaluative family is much bigger than the deontic family and it has much tighter links with affective reactions. In contrast to evaluative concepts, deontic concepts do not admit of degrees. Their logical form is not the same; evaluative concepts, but not deontic concepts, at least apparently correspond to simple predicates. Evaluative concepts are omnivorous, while deontic concepts are concerned with what is at least indirectly subject to the will, or, in the case of *ought to be,* with the state of things. And lastly, value conflicts are not authentic dilemmas; the principle *ought* implies *can* and the principle of agglomeration have no kind of plausibility in the case of evaluative judgements.

3.2 Bridges Between the Normative and the Evaluative

What does all this imply about whether it is possible to accept that evaluative concepts and judgements involving these concepts are normative? The differences that we have examined suggest that the two kinds of concepts serve functions that are too different for it to be reasonable to propose conceptual reductions. Evaluative concepts let us describe and compare different things around us according to a great variety of criteria, corresponding to our diverse affective reactions and allowing for all sorts of nuance. Deontic concepts, on the other hand, concern what we ought or ought not to do, or what ought or ought not to be. There seems to be no reason why we should be tempted to relinquish the services that either kind of concepts provides.[14] But this observation does not resolve the question of whether evaluative concepts are normative. On the contrary, the more it seems that the two kinds of concepts are distinct, the less we can see how they can belong to the same class.

To answer the question of how evaluative concepts can be considered normative, we will have to tackle two tasks that are far from easy. The first consists in determining what makes a concept normative. There are two principal and conflicting conceptions of normativity: The first says that the concept of ought is the central normative concept; and the second that it is the concept of reason or, more precisely, normative reason, that plays this role (for the first conception, see Dancy 2000a, b; and Broome 2004. For the second, see Raz 1999, 2010; Scanlon 1998; Skorupski 2007; Wallace 2010). A concept is normative if it is linked to one or other of these two concepts, depending on which conception is advocated. This link can be considered in the first instance as permitting a reduction to the concept that is normative *par excellence,* whether this is that of ought or that of reason. However, nothing excludes a more liberal position, whereby what counts is simply the ability to establish inferential links. The second task consists in examining all the possible

[14] See Ogien and Tappolet (2009, pp. 121–122), for an argument along the same lines, based on the idea that evaluative considerations give us reasons to act.

links between evaluative concepts and the central normative concept, whether this is that of ought of that of reason. Rather than settling for one of the two conceptions of normativity, I will consider the options available to adherents of each rival view. As will become apparent, there are in fact many inferential links between evaluative concepts, deontic concepts and the concept of reason.

The first option that I would like to discuss assumes that the concept of ought is the central normative concept. The question of the normativity of evaluative concepts would thus reduce to the question of what is the link between evaluative and deontic concepts. Given the distinction between *ought to do* and *ought to be,* we should divide this question into two. Let us first consider the version claiming that *ought to be* is the central normative concept. Evaluative concepts will be normative in as far as they are connected to the concept of ought to be. This is exactly what Jonathan Dancy suggests:

> It is often said that normativity is the characteristic common to everything that appears on the 'ought' side of the distinction between what is and what ought to be. This is true however […] only if we include what is good and bad under the general heading of what ought to be or not to be (Dancy 2000b, p. vii).

The question, evidently, is whether one can count what is good and bad, and more generally all the different ways of being good and bad, as part of the category of what ought to be. To defend this approach, one could argue that, if it is true that something is good, it is true that that thing ought to be.[15] In truth, it does not seem that the fact that something is good is enough to conclude that it ought to be. It rather seems that what ought to be is what is best (Wedgwood 2009, p. 512). Since it also seems plausible to say that if something ought to be, that thing is the best, one obtains the following principle:

1. *x* is the best if and only if *x* ought to be.

An initial question that arises is how to integrate the specific evaluative concepts, such as *courageous* or *admirable*. Possession of such a characteristic, even to the highest degree, does not imply that something ought to be. The most courageous or admirable action is not necessarily the action that ought to be because one cannot exclude the possibility that that action is not the best action—another action could be better, after all. The different specific evaluative characteristics determine if a thing is the best, or more exactly if it is the best all things considered, but specific evaluative judgements do not directly imply judgements about what is the best or what ought to be. Thus, specific evaluative concepts are normative in as far as they contribute to determining the comparative value that something possesses, all things considered.

[15] It is Moore (1966)who argues: 'Every one does in fact understand the question "Is this good?" When he thinks of it, his state of mind is different from what it would be, were he asked "Is this pleasant, or desired, or approved?" It has a distinct meaning for him, even though he may not recognise in what respect it is distinct. Whenever he thinks of "intrinsic value", or "intrinsic worth", or says that a thing "ought to exist", he has before his mind the unique object—the unique property of things—that I mean by "good".' (1903, Sect. 13, 68) See also Mulligan (1989), for the claim that to judge something good implies that that thing should be. Mulligan suggests that the unity of the normative domain is due to the fact that ought to do, like good, implies ought to be.

Another question that this suggestion raises is the extent to which this link with ought to be properly renders the idea that evaluative judgements are normative. What we seem to lose is the link with the idea that normative judgements are judgements that guide our actions. This suggests that it is rather the concept of ought to do that is the normative concept *par excellence*. The difficulty is that even if it is without doubt true that, if an agent ought to perform an action, that action ought to be, ought to be does not seem to directly imply ought to do (see Mulligan 1989, for this suggestion). That world peace ought to exist does not imply anything concerning what particular agents ought to do. After all, it is almost impossible to do anything to contribute to world peace. However, there is a way of skirting around this difficulty by suggesting that we should limit what an agent ought to do to that which she is capable of doing. Thus, one can propose the following principle:

2. *S* ought to φ if and only if *S* is capable of φ and φ ought to be.

This principle allows us to highlight the link between evaluative concepts and the concept *ought to do*. Expressed differently, the principle in question claims that an agent ought to perform the action that is the best among those she is capable of performing:

3. *S* ought to φ if and only if *S* is capable of φ and φ is the best of all actions.

Some will object that this principle implies consequentialism, at the very least a controversial doctrine, and so should be rejected. Indeed, if (3) were a conceptual truth, we would have to conclude that the numerous opponents of consequentialism were not only wrong, but did not properly understand ordinary concepts. What we should note, however, is that it is possible to understand (3) in a non-consequentialist manner. It is sufficient to define what counts as an action that ought to be, or even the best action, in non-consequentialist terms. One can, for example, suggest that what counts is what is good relative to the agent, given the duties that fall to her (see Wedgwood 2009, for this suggestion). From this point of view, the best action for an agent can be not to lie, even if a lie would have the best consequences in neutral terms in the agent's view—she would save more lives, for example.

Furthermore, another link between evaluative concepts and the concept of ought concerns the affective reactions we ought to have towards values. Thus, it seems plausible to say that we ought to approve of what is good, disapprove of what is bad, admire what is admirable, despise what is despicable, etc. This is one of the interpretations of the idea that value concepts can be elucidated in terms of what are called appropriate (or fitting) reactions (Brentano 1889; Wiggins 1987; Mulligan 1998; Scanlon 1998; D'Arms and Jacobson 2000, among others). More generally, we have:

4. *x* is *V* if and only if *x* is such that *S* ought to *R* towards *x* (where '*V*' is an evaluative predicate and '*R*' is an affective reaction towards *S*).

The question of how exactly to formulate this kind of equivalence remains tricky. For example, we can ask how we should understand the term 'ought'.[16] However, it

[16] In certain uses of the term, the equivalence is clearly false. Something can be amusing, even if from a moral point of view one ought not be amused (D'Arms and Jacobson 2000; Rabinowicz and Rønnow-Rasmussen 2004).

is difficult to deny the plausibility of such an equivalence, which makes it plausible that a formulation that makes it true exists. Furthermore, even if the equivalences are often proposed with the aim of reducing evaluative concepts to other kinds of concept, in this case to deontic concepts involving our reactions, this is not the only possibility. We can think that what such equivalence shows is that there is a tight connection between the two kinds of concepts, without this implying an asymmetry (Wedgwood 2009).[17]

Another formulation of the idea that there is a link between value judgements and our reactions uses the concept of reason, rather than the concept of ought (see Scanlon 1998, for example). According to Thomas Scanlon, evaluative judgements are not only linked to judgements involving our affective reactions, but also to our practical judgements. More precisely, Scanlon claims that something is good in as far as it possesses the natural properties that give us reasons to act or to react positively towards that thing. For Scanlon, the thought is that showing that something is good is nothing more than showing that it possesses the traits which provide reasons (for a critical discussion, see Ogien and Tappolet 2009, Chap. 3). However, a reductionist reading is not the only reading here. One can subscribe to the idea that if a thing gives reasons to do something or to feel something, it is precisely because it possesses value. To leave both possibilities open, one can simply propose the following equivalence:

5. *x* is good if and only if *x* gives a reason to perform certain actions and to have *R* towards *x*.

This leaves one free to say that, if something provides reasons, this is simply in virtue of its natural properties. This claim is as compatible with (5) as the claim that reasons is based on the evaluative properties of things.

In any case, this equivalence, which one cannot deny is plausible, allows us to render the normative character of evaluative concepts within the framework of a conception that states that the normative concept *par excellence* is that of reason. A point worth underlining is that this conception allows us—and more directly than the conception that privileges the concept of ought—to render the normative character of specific evaluative concepts. Indeed, (5) can be formulated for specific evaluative concepts just as well as it can for general evaluative concepts. It seems plausible that something is admirable as far as it gives us reasons to act, and overall to feel admiration towards it.

However, we should keep in mind that the concept of reason and the concept of ought are also connected. Few would deny that we ought to perform an action if and only if we have sufficient reason to do it. Indeed, this is a claim that can be as easily accepted by someone who privileges the concept of ought as by someone who privileges the concept of reason.[18] Following on from this, it is possible to claim that, as far as the fact that something possesses such and such a value gives us a reason to

[17] Also see Tappolet (2011), for the claim that, if it is true that one ought to feel such and such a reaction in response to something, it is because we want to have correct reactions in response to things, where *correct* is not a normative concept.

[18] For a version of the claim that privileges the concept of ought (*ought-first*), Broome (2004, p. 24 and 39). According to Broome, this equivalence is not analytic, but is implied by the fact that the reason for doing something is an explanation of why one ought to do that thing.

act, the fact of possessing a value is linked to what we ought to do. In giving us reasons to act, values contribute to determining what we ought to do. The upshot is that it is not surprising at all that values and their concepts are considered normative.

The picture that crystallises is one in which a great many equivalences allow us to build bridges between the evaluative and deontic domains. Doubtless, we need to formulate these equivalences in a more precise manner. However, it is difficult to deny their plausibility. One might thus think that there is only a little work needed to show that all these different concepts can be reduced to each other. If the only concepts we needed were deontic, for example, we would lose nothing if, suddenly, from one day to the next, we stopped using evaluative concepts.

I think that the conception that emerges is rather different. On the contrary, what the existence of multiple links suggests is that evaluative concepts and deontic concepts are two kinds of concepts that belong to the same conceptual level. Neither one nor the other of the two families should be considered prior. As Wedgwood claims, these concepts are too closely linked for it to be plausible to say that some have conceptual priority over the others (Wedgwood 2009, p. 513).[19] Even if I have not shown that this conception is inevitable, I think one must concede that it is not only possible, but attractive.

3.3 Conclusion

In brief, the reply to the question of whether evaluative concepts can be considered normative is the following: They can because they possess a great number of inferential relations with both deontic concepts and the concept of reason. The normative domain, although made up of many different kinds of concepts, is a unified domain.

It should be clear that this way of conceiving of the normative domain corresponds to the image that our examination of the distinctions has provided us with. Indeed, as we have seen, the differences between evaluative and deontic concepts suggests that the two kinds of concepts fulfil distinct functions: For evaluative concepts, this consists in the description and comparison of things around us, including people and their actions, according to a variety of criteria and nuances corresponding to multiple affective reactions; for deontic concepts, this consists in a verdict on what one ought or ought not to do, or on what ought or ought not to be.

Acknowledgments My deepest gratitude is to Kevin Mulligan, for introducing me to a great many philosophical distinctions, among which that between the evaluative and the deontic, when he supervised my PhD thesis. The first version of this chapter has been written for the 2010 Hughes Leblanc Conferences featuring Kevin Mulligan, and organised by Denis Fisette, whom I wish to thank. I also would like to thank Daniel Laurier, Jonas Olson, Francén Ragnar and Claude Panaccio for their questions and comments. A French version of the chapter has appeared in *Philosophiques* (2011, 38(1):157–76). Many thanks to Chloë Fitzgerald for the English translation. My work for this chapter was supported by research grants from the FQRSC and the SSHRC, which I gratefully acknowledge.

[19] Wedgwood suggests that the case is roughly comparable to the relation between possible and necessary.

References

Brentano FC (1889) Vom Ursprung sittlicher Erkenntnis, 1st edn. Felix Meiner, Leipzig (1955)
Broome J (2004) Reasons. In: Wallace RJ, Pettit P, Smith M, Scheffler S (eds) Reason and value. Themes from the philosophy of Joseph Raz. Oxford University Press, New York
Bykvist K (2009) No good fit: why the fitting attitude analysis fails. Mind 118:1–30
Carnap R (1935) Philosophy and logical syntax. Kegan Paul, Trench, Trubner & Co., London
Cuneo T (2007) The normative web. An argument for moral realism. Oxford University Press, Oxford
Dancy J (2000a) Practical reality. Oxford University Press, Oxford
Dancy J (2000b) Editors's introduction. In: Dancy J (ed) Normativity. Blackwell, Oxford
D'Arms J, Jacobson D (2000) Sentiment and value. Ethics 110:722–748
Geach PT (1956) Good and Evil. Analysis 17:32–42
Geach PT (1991) Whatever happened to deontic logic? In Geach P (ed) Logic and ethics. Kluwer, Dordrecht
Hansson SO (2001) The structure of values and norms. Cambridge University Press, Cambridge
Hare RM (1952) The language of morals. Clarendon Press, Oxford
Hume D (1739–1740) A treatise of human nature, 2nd edn. Oxford University Press, Oxford
Heyd D (1982) Supererogation. Cambridge University Press, Cambridge
McNamara P (2006) Deontic Logic. In: Zalta EN (ed) The Stanford encyclopedia of philosophy. http://plato.stanford.edu/entries/logic-deontic. Revised 2010. Accessed 1 Nov 2013
Moore GE (1903) Principia Ethica. Cambridge University Press, Cambridge (revised edition, 1993)
Mulligan K (1989) Wie verhalten sich Normen und Werte Zueinander? Manuscript
Mulligan K (1998) From appropriate emotions to values. The Monist 81(1):161–88
Mulligan K (2009) Values. In: Le Poidevin R, Simons P, McGonigal A, Cameron R (eds) The Routledge companion to metaphysics. Routledge, London
Ogien R, Tappolet C (2009) Les concepts de l'éthique. Faut-il être conséquentialiste? Hermann, Paris
Rabinowicz W, Rønnow-Rasmussen T (2004) The strike of the demon: on fitting pro-attitudes and value. Ethics 114/3:391–423
Raz J (1999) Engaging reason. Oxford University Press, Oxford
Raz J (2010) Reason, reasons and normativity. In: Shafer-Landau R (ed) Oxford Studies in Meta-ethics, vol. 5 . Oxford University Press, New York
Ross WD (1930) The right and the good. Oxford University Press, Oxford
Scanlon TM (1998) What we owe to each other. Harvard University Press, Harvard
Schroeder M (2008) Value theory. In: Zalta EN (ed) The Stanford encyclopedia of philosophy. http://plato.stanford.edu/entries/value-theory/. Accessed 8 Dec 2011
Schroeder M (2011) Ought, agents, and actions. Philos Rev 120(1):1–41
Skorupski J (2007) What is normativity? Disputatio 2(23):247–269
Smith M (2005) Meta-ethics. In: Jackson F, Smith M (eds) The Oxford handbook of contemporary philosophy. Oxford University Press, Oxford
Thomson JJ (1992) On some ways in which a thing can be good. Soc Philos Policy 2:96–117
Thomson JJ (2007) The right and the good. J Philos 94:273–298
Thomson JJ (2008) Normativity. Open Court, Peru
Tappolet C (2004) Les dilemmes moraux et les devoirs prima facie. In: Canto-Sperber M (ed) Dictionnaire d'éthique de philosophie morale, 4th edn. Presses Universitaires de France, Paris
Tappolet C (2011) Values and emotions: the prospects of neo-sentimentalism. In: Bagnoli C (ed) Morality and the emotions. Oxford University Press, Oxford
Väyrynen P (2010) A wrong turn to reason. In: Brady M (ed) New waves in metaethics. Palgrave Macmillan, New York
von Wright GH (1963) The varieties of goodness. Routledge and Kegan Paul, London
Wallace RJ (2010) Reasons, values and agent-relativity. Dialectica 64(4):503–528

Wedgwood R (2009) The “good” and the “right” revisited. Philos Perspect 23:499–519
Wiggins D (1976) Truth, Invention and the Meaning of Life. In: Needs, Values, Truth. Blackwell, Oxford
Wiggins D (1987) A sensible subjectivism. In: Wiggins D (ed) Needs, values, truth: essays in the philosophy of value. Basil Blackwell, Oxford
Williams B (1965) Ethical consistency. Proc Aristot Soc 39:103–124
Williams B (1985) Ethics and the Limits of philosophy. Fontana, London

Chapter 4
For Kevin's Sake

Toni Rønnow-Rasmussen

Abstract The idiom 'for someone's sake' plays a central role in recent attempts to understand the distinction between impersonal values and personal values—e.g. between what is valuable or good, period, and what is valuable *for* or good *for someone*. In the first section, three historical approaches to this distinction are outlined. Sect. 4.2 presents a modified fitting-attitude (FA) analysis of final 'value-for' interpreting value-for in terms of there being a reason to favour something 'for someone's sake'. Sect. 4.3 outlines two arguments against this sort of modified analysis, and then indicates what the rejection of these arguments would involve. This section also identifies an ambiguity in the analysis deriving from the fact that 'sake' may be used either evaluatively or nonevaluatively (descriptively). In Sect. 4.4, the modified FA analysis is further clarified. Sect. 4.5 focuses on Kevin Mulligan's recent suggestion that we are struck by personal value; finally, in Sect. 4.6, it is shown that an FA analysis admitting of two varieties of goodness may help us understand a certain kind of case that appears paradoxical as long as we assume that there is good, period, and no good-for.

Keywords For someone's sake · Fitting-attitude analysis · Good-for · Personal value · Distance problem

Moral philosophy and value theory in general frequently make use of the idea that we should do something for someone's sake. Notwithstanding this, 'doing something for someone's sake' remains a somewhat obscure notion. Of course, examples of reasons or motives for doing such acts easily come to mind. For instance, I could come up with a number of reasons why I have contributed to this festschrift. Some relate to what I believe would be valuable for me. My chapter should strengthen my association with a great philosopher, one from whom I have learned a lot over the years and hope to learn more—unless, of course, with this contribution, I have made a mockery of myself! And that association will very probably be good for me. But whatever purely prudential motives I might have, I am clearly motivated by the fact that he is a very good-hearted and amiable fellow whose emblematic Mulligian wit

T. Rønnow-Rasmussen (✉)
Lund University, Lund, Sweden
e-mail: toni.ronnow-rasmussen@fil.lu.se

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_4,

has made me burst into laughter on many an occasion. So, given who Kevin Mulligan is, I would say I have strong reasons to contribute to this collective tribute for his sake, not just my own. In fact, in my view there is a positive personal value accruing to this homage—a value for Kevin.[1] This personal value is best understood, I think, in terms of there being a reason to respond in a positive attitudinal way to this festschrift, and to do so for Kevin's sake.

But although I have some idea of how to answer our original question, I am still not quite sure what, precisely, it means to do something for someone's sake. Why am I doing this for Kevin's sake rather than for, say, Kevin? What, *if anything,* is the difference? What exactly is this *sake* that apparently plays some motivational role in my deliberation? These questions interest me for more than one reason. Besides the obvious one (that I want to know, of course, what I am doing), I have also a special interest in clarifying matters. Recently I examined in Rønnow-Rasmussen (2011) the possibility of expanding our value taxonomy with a new set of values. In doing so, I argued that there seemingly are two kinds of nonderivative value: value, period, and value-for. For instance, if we have a thin, fundamental, positive value in mind, such as goodness, there is, I argued, a conceptual role to be played not only by good, period, but also by the equally thin, fundamental, nonderivative good-for.

To understand these sorts of value, I applied a slightly modified variety of an analysis which has, in recent years, attracted a good deal of attention among value theorists. On this analysis—the so-called fitting-attitude (FA) analysis of value[2]—final values are understood in term of attitudes that it is fitting, or, more generally, that there is reason, to direct towards the value-bearer for its own sake. 'For someone/something's sake' attitudes obviously play a central role in this analysis. The kind of attitude involved in the analysis of value-for is more complex than the one required for the analysis of value, period.

In what follows, I will consider an objection to my analysis which sets out from the idea that the idiom 'for someone's sake' is not obviously translatable into other languages. Besides allowing me to once more determine the role of the noun 'sake' in the analysis, the discussion will, I hope, highlight the fact that there are at least two ways to understand the notion of doing something for Kevin's sake. In addition to killing two birds with one stone, I will also, in the final section, outline why I think it might be a good idea for an FA analyst to allow that there are two kinds of value: impersonal and personal. Meanwhile, since there is no consensus on this matter, I will begin by briefly saying something in support of the very idea that this sort of distinction can be made.

[1] For the rest of the chapter, I will refer to the 'Jubilar' in a more friendly way as Kevin.

[2] Following Scanlon (1998), one version of the FA analysis is called the 'buck-passing' account.

4.1 Monism and Pluralism About Varieties of Good

Here is an example in which I feel that the distinction between good and good-for easily comes into play: The World Chess Championship 1972, which took place in Reykjavik, is considered by many aficionados as *the* match of the twentieth century. Bobby Fisher, of the USA, beat the defending champion, Boris Spassky, of the Soviet Union. It is sometimes said that, in making various unexpected demands and being, in general, unreasonable about the whole setting of the match, the eccentric Fisher was using mind games to undermine Spassky's confidence. Whether this is true is a matter of interpretation that I am not competent to get into (I was not there). However, it would be quite understandable for those who believe that Fisher 'psyched out' Spassky to think both that accomplishing this was good for Fisher and bad for Spassky *and* that it was bad, period, that the match was settled in an unfair way. Fairness and justice are bona fide examples of value, period.

Somewhat surprisingly, the idea that we should recognize good-for as another kind of value, alongside good, period, has not been generally accepted. In fact, it is possible to detect three major positions with regard to the distinction between good and good-for. Moorean monists, as I have referred to them elsewhere (Rønnow-Rasmussen 2013), believe that there is no good-for—or at most that this kind of goodness is somehow derivative from good, period. Moorean monists claim there is just one, unique kind of fundamental positive value (goodness), and that this is a nonrelative, absolute or final. This is the kind of goodness that I am referring to as good, period. Of course, Mooreans do not deny that we sometimes speak about something being good for someone. However, they insist that, when this happens, the good-for is reducible to, or definable in terms of, something's contribution to, or constitutive role in, what is good, period—e.g. the goodness of person's life. Thus, the claim that Fisher's behaviour was good for Fisher should be understood to mean that there is some good, period, accruing to a state of affairs such as the one involving Fisher's state of mind when he checkmated Spassky.

For Hobbesian monists—as I will refer to them, given Hobbes' insistence that there is nothing that is simply good, only 'goodness to us'[3]—whenever we speak about goodness, it will always be because something is good for someone. There are reasons why one would be attracted by this sort of sinister claim (besides, I suppose, its *sinisterness,* which does seem to attract some people). But, fundamentally, Hobbesians will not accept the Moorean attempt to understand what is good for Fisher in terms of a good, period. Intuitively, it does seem reasonable to describe the case as involving something that is good for one person, and something that is bad for another person. It is less easy to see how one could be led to believe that all value is somehow person-relative in this sense.[4] For instance, Mooreans will be eager to point out that the badness of injustice should not be regarded as someone's badness (unless, of course, this person-relative badness is at some point reducible

[3] I strongly suspect (but admit that it is a matter of interpretation) that Hobbes himself was a Hobbesian monist (e.g. Hobbes (1969/1889), Chap. 7, p. 29).

[4] Recall Prichard's (1928, pp. 21–49) point that Plato regarded even justice as a kind of good-for.

to badness, period). The Mooreans intuitively seem to have a point; but, again, it seems that they too are guilty of a adopting a narrow perspective.

Neither of these positions, in my view, is convincing. A more compelling view is taken by value pluralists, and specifically dualists, who recognize both kinds of value—the relative as well as the nonrelative good—and conceive of these as phenomena that are understandable independently of each other (which, by the way, is quite consistent with the notion that both are analysable with the same pattern of analysis). Value pluralism squares better with our intuitions; it offers a straightforward approach to the question how we should look upon the values involved in cases like the chess example. Both kinds of monist will have more explaining to do here. Moorean monists will try to do so, typically, by arguing that it does not make sense to say that an event, object or state of affairs is good for a person unless that event, object or state of affairs constitutes, or at least contributes to, the *impersonal* goodness of something else, such as the good, period, that they take to accrue to the life, welfare or wellbeing of the person. Not only dualists, but also the Hobbesian monist, will contest this claim, though. One line of reasoning is that there is no need to bring in any value-claim other than that something carries a value-for. Things that carry value-for often also carry value, period. For instance, the present festschrift contains contributions all of which (with the possible exception of the present one) have value independently of their relation to Kevin. So it is expected that the festschrift will carry value, period, as well as a value for Kevin. Non-Mooreans would not deny this possibility; but they would insist as well that there is no conceptual need to admit that something like a festschrift has value, period, only because it has a positive value for a person (for more on the monist/pluralism matter, see Rønnow-Rasmussen (2013).

Naturally, value pluralism is the more complex view; it makes value aggregation and comparison quite complicated, and perhaps even impossible. Here, monists do have an advantage over dualists. However, this is not the place to settle the controversy between monists and pluralists. I bring it up because I want to stress that some of the problems I have run into, in understanding good-for in terms of someone's-sake attitudes, are connected with my belief that we should not make this analysis dependent on ascribing value, period, to the person or, for that matter, his or her sake.

4.2 Analyzing Value-For

According to the sort of view I am trying to understand, value-for might well accrue to an object rather than the person whose personal value it is. Again, if we have a positive value like goodness in mind, it might well be that 'this festschrift carries *value for* Kevin' should be understood to be about a person-relative value[5] that

[5] The expression 'person-relative value' is ambiguous. It might mean that the value is constituted by a certain subject (as a subjectivist would say that is the case with all values), or that it is a value

accrues to a certain festschrift. Suppose this makes sense. Now, since such an object might well also carry impersonal value, an FA analysis has to tell us wherein lies the difference. In *Personal Value,* I came to the conclusion that the distinction should be located, not in the normative component, but rather in the attitudinal element.[6] In the case an object carries both kinds of final value, given the analysis, there is a reason to favour the object for its own sake. It is just that when it comes to personal, final value, the sort of attitude involved is more complex: It is a case of favouring something for its own sake for the sake of someone else. We must also take into account that, just as there can be final and instrumental value, there can be final and instrumental value-for. However, here I will confine myself to so-called *final* personal value. In the case of *positive* final personal value, i.e. value-for, I suggested the following **F**itting-attitude **A**nalysis of **P**ersonal value:

> FAP: an object *x's value for* a person a consists in the existence of normative reasons for favouring *x* for its own sake for *a*'s sake

'Favouring *x* for its own sake for *a's* sake' is to be understood as a schematic placeholder for the different pro-responses that are called for by different kinds of valuable object. I devoted considerable space in *Personal Value* to the clarification of linguistically awkward-sounding attitudes of this sort, which I referred to as final for-someone's-sake attitudes ('FFS-attitudes'). It would require too much space here to repeat my responses to the various questions and objections that I considered. Instead, I would like to consider a possible objection to FAP that has to do with the attitude involved—one that I did not discuss in *Personal Value*.

The objection is quite radical.[7] It deserves perhaps to be carved out in more detail than I can manage here, but the general idea is easy to state. It comes down to the claim that FAP should be rejected because the notion of an FFS attitude appears to be translatable only in some languages, and we simply cannot, the argument goes, tolerate an analysans that fails to work in some languages.

4.3 The Non-Translatability Objection

The gist of what we might suitably call the non-translatability objection can be straightforwardly stated. An analysis aspiring, besides being true, to further conceptual clarity should not couch its analysans in terms that are not translatable into other major natural languages. Just how reasonable this claim is depends in part on

---

from a certain person's perspective. However, this is not how it is intended to be understood here. Rather, if something is good for a person, this something carries a person-relative value, but this sort of value is not necessarily only analysable by subjectivists; nor is it goodness only from a certain perspective.

[6] Nor is this necessarily a distinction between an 'objectivist' and a 'subjectivist' approach to value.

[7] I would like to stress that there are a number of other reasonable and interesting objections to FAP, but that I will not consider them here.

what we understand by ‘translatable’. For instance, we might require that the statement in the analysans should be fully translatable. By ‘fully translatable’ I mean that the translation should contain word-for-word equivalents of the original statement. But such a claim would surely be exaggerated. Translations from one natural language to another are hardly ever of this exact kind. A more reasonable idea is to require, not a one-for-one replacement, but rather that the translation should express the core meaning of the original statement. Just what this means, other than that the translation should express to a high degree what the original statement expresses, is a difficult question. However, this question need not, I believe, detain us here. Obviously, ‘core meaning’ is open to different kinds of qualification.

So let us rephrase the objection, taking this more cautious claim into account:

The non-translatability objection:

1. An analysis in one natural language should be rejected if the core meaning of the analysans cannot be expressed by another natural language
2. FAP is formulated in terms of the expression ‘for someone’s sake’
3. Some natural languages have no one-for-one replacement of ‘for someone’s sake
4. Natural languages that do not contain a one-for-one replacement of ‘for someone’s sake’ cannot express the core meaning of the FAP’s analysans
5. FAP should therefore be rejected; it cannot be expressed in some natural languages

First, although this modification seems justified, it should be stressed that the kind of objection is not obviously reasonable for analyses in general. At least it seems wise to keep the door open to the possibility that, at some point in time, some natural languages might be conceptually more evolved or fine grained than other natural languages. If that were the case, there would not necessarily be anything wrong with an analysis just because, at some point in time, it resists translation into another language, i.e. cannot be exactly expressed in that language. Still, although this might be the case with certain terms and expressions, it is hard to see that this comment really applies to our case. FAP is about good-for (value-for). Some of the languages apparently lacking a one-for-one replacement for ‘for someone’s sake’ contain words expressing the idea that something is good, or valuable, for an agent. It would be very odd, from the point of view of the analysis, if we could not find a translation capturing approximately the same meaning even if that were done without any synonym of ‘sake’ as it is used in ‘for his sake’.[8]

Here is a list of the ways in which ‘for his sake’ would ordinarily be translated into some languages (most of which Kevin has mastered):

Catalan: Pel seu bé
German: Für sein Willen
French: Pour son bien
Italian: Per il suo bene
Spanish: Por su bien
Swedish: För hans skull

[8] The objection rests on several notions and ideas that stand in need of clarification, and so it is quite likely that it can be criticized for other reasons. For the sake of advancing the discussion, I have set these criticisms aside.

A couple of features of this short list are worth pointing out. First, some languages translate the expression into what philosophers, at least, would describe as evaluative language. For example, 'bé' in Catalan, 'bien' in French, 'bene' in Italian, and 'bien' in Spanish, translate the English word 'good'. However, translators in other languages seem to have thought they could do without an explicit 'evaluative translation'. Thus the Swedish 'skull' probably would not be straightforwardly classified as an evaluative term. The same can be said of the German 'willen', and, of course, the English 'sake'. (That they admit of evaluative interpretations is another matter, to which I will return in a moment.) In the Latin languages, the point is even more obvious, since they think that, in some contexts, they can manage without even a synonym of 'sake'. For instance, in French 'pour lui' would in some contexts work as translation of 'for his sake'.

The list I have provided is admittedly too short and culturally myopic. Nonetheless it raises some interesting issues[9] and provides evidence that 'sake' is an ambiguous term—one that may or may not express an evaluative idea. This conclusion is strengthened once we take into consideration a further aspect. Some languages appear to have more than one way of translating 'for his sake'; and, interestingly, both evaluative and nonevaluative translations exist within one and the same language. For instance, the distinction between 'para' and 'por' in Spanish is food for thought. 'Para' and 'por' signal quite different things. If you are employed by your son, you might express this in Spanish as 'Trabajo *para* mi hijo'. However, if you work for your son's sake you might say: 'Trabajo *por* mi hijo'. Sentences combining 'para' and 'por', aimed to express this distinction, often appear strange. But perhaps this would work. Someone taking care of a very noisy parrot, not because he cares about the bird, but as a favour to its holidaying owner, might say: 'I bought food for the bird for Franscisco's sake'. Spanish might express this in different ways, but one way would be to translate this as: 'Compré comida *para* el pajaro *por* Franscisco'.

The 'para'/'por' distinction is important if we wish to understand the nonevaluative notion of 'sake'. Clearly, it signals that there are at least two ways in which we can do something *for* someone (say, the bird or Franscisco). We can do it in the 'para' or the 'por' way. That is, I do not want to deny that there is more than one sense involved when we speak about doing something 'para'/'por' someone (i.e. that the distinction fulfils more than one function in language). However, among these senses, I believe there is, for my purpose, one salient one. Both 'para' and 'por' refer, in a sense, to the cause, motive or reason-constitutive ground[10] that explains why you did (or are doing, or will do) something, i.e. the purpose for which you act. However, using 'por' signals that we should not look for any further purpose. The thing, or the person, we did it 'for'/'por' is the final end—the cause, motive or reason for our acting or favouring. This is not the case with 'para'. Saying

[9] I do not have in mind, primarily at any rate, questions about the value-theoretical background of translators: But perhaps it would be an idea to introduce some meta-ethics and formal axiology in the translator curriculum?

[10] It should be borne in mind that 'sake' expressions are customarily translated into Latin by 'causa' or 'gratia'.

that we did it 'for'/'para' someone leaves at least the ultimate reason, cause or motive for which you did something open.

Languages have different ways of expressing this distinction. Although it is an entirely different kind of word, the Swedish 'skull' plays a role that is similar to 'por' in one respect.[11] When a Swedish speaker says that he has no further reason for doing something than that already referring to some person or wants to be explicit about not having an ulterior motive, he may do so by saying that he did it for this person's 'skull' (i.e. 'sake'). 'Skull' may refer to no specific set of properties of the person other than, say, 'being Kevin'. Of course, we might, and I suspect often do, have a precise set of properties in mind—e.g. the properties that constitute Kevin's welfare, wellbeing, or whatever is in his interest.[12]

I am no expert in French, but my impression is that it often accomplishes the same sort of distinction by iteration. For instance, 'art for art's sake', is frequently translated as 'l'art pour l'art'', and when, for instance, Barbara chants, at the beginning of her song,

> A mourir pour mourir,
> Je choisis l'âge tendre,
> Et partir pour partir
> Je ne veux pas attendre,
> Je ne veux pas attendre,

we feel (perhaps more than understand) that she wants to die for dying's sake, and to leave for the sake of leaving. So I venture to say that neither French nor Spanish has a problem expressing the core meaning of an analysans formulated in terms of 'sake', even when we take 'sake' in its nonevaluative sense. Naturally, this is but an outline of an argument the details of which will have to be left to another time. But in view of what I have said so far it is plausible to reject premise (4) of the non-translatability objection. Languages that do not employ the 'sake' idiom can still express what FAP states. For instance, languages that we have considered here have different means of expressing what FAP expresses with 'sake'.[13]

It might be said that even if the non-translatability objection is not much of an objection, some of the translations we considered introduce another difficulty. When FAP is understood in terms of the nonevaluative 'sake' it looks pleonastic; it employs one word too many to express its analysans; 'sake' is not needed. However, as we have seen, this objection is strictly not true. Simple removal of 'sake' from the analysans in FAP would strand us with a less clear formulation. Thus, saying 'I am doing something for *x*' and saying 'I am doing something for *x*'s sake' might

[11] Etymologically 'skull' is related to 'or*sak*' (cause) as well as to 'skuld' (culpability). This is even clearer in Danish, which translates our phrase as follows: 'For hans *skyld*'; 'being guilty' becomes in Swedish, as well as in Danish, being 'skyldig'.

[12] The complex relationship between good-for and notions such as welfare and wellbeing is examined in Rønnow-Rasmussen (2011).

[13] Understanding final values in terms of 'sake' is not obviously right, though. Another possibility would perhaps be to have an analysis in terms of 'favouring something as an end'. However, there are different disadvantages with this suggestion (Rabinowicz and Rønnow-Rasmussen 2000, pp. 47–48).

well come close to the same meaning. However, it need not always be the case. The latter does not leave it open, as the former does, whether your act is being performed with an ulterior motive. The former makes it quite clear that the final 'end' of your act refers to *x*. To modify FAP as suggested above would therefore make it ambiguous in a way that it is not in its original form. Hence, this objection is not really convincing.

4.4 Clarifying FAP

However, as we have seen, 'sake' is in another sense itself ambiguous, since it can be understood in an evaluative or a nonevaluative way. This gives rise to the further question whether FAP should be interpreted as involving the evaluative or the nonevaluative 'sake'? Should we, in other words, understand personal values in terms of reasons for favouring something 'for someone's good/bé/bien/bene'? Or should we take 'sake' in its nonevaluative sense—a sense signalling that the purpose, or the reason why we should favour something, is finally located in the person, period (for him or her/for his or her 'skull'/skyld/Willen)?

From the point of view of FA analysis the response seems quite obvious. If values are to be understood, ultimately, in terms of the normative, the evaluative interpretation is not, unless it is qualified, at all welcome. We do not want to run into a value in the analysans if we can avoid it. In fact, even for the FA sceptic there is reason to be cautious about including a value in the analysans. One need not deny that sometimes, when we judge that we have reason to favour something for someone's sake, what we have in mind is favouring for this person's good. I am sure this occurs regularly. Indeed I think it is quite understandable that the sort of things that we regard as carrying personal value are precisely things that relate to people we somehow estimate, i.e. think of in terms of some value notion. Recall, for instance, what I said at the outset. When I discussed the value this festschrift has for Kevin and the rest of us, I mentioned several valuable features of Kevin. But notice, acknowledging this does not imply that the analysis of personal value must be couched in terms of some value accruing to the person. So although we are as a matter of fact interested in the personal value of people that we hold in high esteem, it would be a mistake to conclude from this that what is positively valuable for a person *a* must be understood in terms of *a's* having a certain positive value. Even bad persons have personal values, be they good or bad for them.

One more thing needs to be added to this picture. In cases in which we ascribe value-for in terms of an evaluative 'sake' it is still an open issue whether this person's good should be understood in terms of a Moorean impersonal good, period, or a Hobbesian good-for. Now, whether or not you adopt FA it would seem that you have at least a prima facie reason to be sceptical about including an impersonal value in an analysis of a personal value. Something simply gets lost in this reduction—namely, the idea that the value is personal. But perhaps such a reduction could nonetheless work. A number of people seem to think so. Moore (1993/1903)

is not alone on this matter. However, for those, like me, who are unconvinced by these attempts, placing 'good, period' in the analysis of 'good-for' appears to be a bad idea.

The Hobbesian alternative suggested that favouring something for someone's sake is, in effect, favouring it for what is good for the person. Now, for obvious reasons, we do not want to analyse, say, *good-for* in terms of *good-for* (unless, perhaps, we qualify these notions in different ways). However, here it seems wise to make haste slowly. Personal values encompass more than what is good for persons. So it might be that some *values-for* should actually be understood in terms of *good-for*.[14] We should, I think (although it is not something I will argue for here), recognize the possibility that some personal values accrue to something in virtue of some other personal value. If that is the case, FAP should be able to handle such cases, too. I do not think it has any problems doing so. All of this would, at least, be consistent with the idea that among reason-constitutive properties there are value properties which themselves have to be analysed in terms of FAP. If there is something to these speculations, we should concede that while some values-for are derivative and non-fundamental, others are nonderivative and fundamental. I see no reason why FAP cannot handle both derivative and nonderivative value-for.

4.5 Vocation and Being Struck by Personal Value

FA analyses face some tough challenges. The most serious is perhaps the so-called wrong kind of reason problem. This has certainly attracted most attention. However, there are other problems. The reductive nature of this analysis squares less well with some of our intuitions about values. For instance, Kevin himself has recently argued in 'On Being Struck by Value—Exclamations, Motivations and Vocations' (see also Mulligan 2009a) that we might be struck by personal value. Here a caveat is in place, however. What Kevin has in mind by 'personal values' is not quite what I am trying to capture with FAP. There is some overlap, though.

Kevin explains what it is like to be struck by personal value as follows:

> We sometimes discover what is non-extrinsically valuable for us, our personal values, what used to be called our vocation (Bestimmung). A vocation has all the properties which, Williams thinks(1981, p. 130), are possessed by what recognition of practical necessity implies: 'an understanding at once of one's powers and incapacities, and of what the world permits, and the recognition of a limit which is neither simply external to the self, nor yet a product of the will'(Williams 1981, pp. 130–131). (Mulligan 2009b, pp. 147–148)

He then makes two important observations.

> Talk about knowledge of one's vocation(s) is doubly ambiguous. First, such knowledge may be negative or positive. The clearest and least controversial cases of discovery in this

[14] For example, what is *desirable for* or *commendable for* me should perhaps be understood in terms of what there is reason to desire or commend with an eye to what is good for me. Perhaps this is an example. It certainly needs to be further examined.

> area are discoveries that a certain way of life is not *for me,* that a certain persona or occupation or habit is *not for me*. Perhaps all such discoveries are negative.[15] Secondly, when we talk about what is intrinsically valuable for Sam or Maria it is easy to overlook the difference between the individual or personal values themselves and their terms. Suppose that certain very specific ways of being generous are intrinsically valuable for Sam. Such individual or personal values are perhaps exemplified by Maria but they could be exemplified by someone else. If it makes any sense at all to talk of a person's vocation, then what is constitutive of such a vocation are the personal values themselves and only secondarily their contingent exemplifications and non-exemplifications. (Mulligan 2009b, p. 148)

Perhaps Kevin would accept that not all personal values are constitutive of a person's vocation. In that case, my notion of a personal value would come closer to his. As we have seen, Kevin also examines the possibility that we are struck by our personal values. I am not sure whether he also would be ready to say that we may be struck by other people's personal value. Perhaps what makes personal values personal is rather the fact that they are constitutive of a person's vocation. Or perhaps it is necessarily the case that you must be struck by the value, *and* that it must relate in the relevant way to your vocation. These later suggestions sit less well with what I have in mind with personal values. In fact, as I see it, a personal value may accrue to some object despite the fact that the person for whom this is a personal value never will be able to recognize this value. Be that as it may. It is a fundamental, substantial issue. Perhaps the more obvious difference has to do with the sort of intuitionism that Kevin also considers—namely, the idea that we are *struck* by this sort of value. As he rightly points out (2009b, p. 159), this challenges a value-reductive analysis such as the FA analysis. For obvious reasons, value experiences do not fit an account that purports to 'do away' with values.

I think Kevin is certainly right about us having what can, in a sense, be characterized as value experiences. However, there are, as Kevin knows better than I, different ways of accounting for such experiences. We should therefore, I think, recognize the possibility that such experiences need not actually involve values in a realistic sense, and hence that the FA analysis need not be jeopardized by value phenomenology.

4.6 A Dualist Dealing With the Distance Problem

FA analysis has more recently been put to another kind of test. Consider the following case. Two children, *A* and *B,* are drowning, and only one of them can be saved by a man in a boat. Now, the state of affairs involving child *A* being saved has the same degree of value as the state of affairs of child *B* being saved. But, suppose now that the man in the boat is the father of *A,* and *B* is someone who he does not know at all. Intuitively, we would think it is only fitting that the father has a more intense desire to save *A* than he has to save *B*. But cashed out in terms of the FA analysis, the

[15] The quote has here the following footnote: 'On this idea in Scheler and Musil, cf. Mulligan "Selbstliebe, Sympathie, Egoismus"'.

latter intuition then tells us that saving *A* is better than saving *B*. This runs counter to the original idea that the same degree of value accrues to saving *A* and to saving *B*.

Krister Bykvist, in a recent paper on the FA analysis, suggests that this sort of case sets a real problem for FA analysis.[16] Basically, the problem is the following. Attitudes may differ in degrees of intensity. But values, too, come in degrees. As we saw, it is fitting that the father in our example wants to save his child with greater intensity. But, again, saving *A* has the same degree of value as saving *B*. What Bykvist argues is that FA analyses cannot handle 'cases where the *degree* to which we should favour (disfavour) something does not correspond to the *degree* to which the thing is good (bad) in itself' (Bykvist 2009, p. 2).[17]

Bykvist also detects the source of this problem.[18] What determines the intensity of the FA is, at least in part, the 'distance' (relation) the 'attitude holder' is from the valued state of affairs. In cases like the one we have considered

> […] the degree to which it is fitting to positively respond to a state of affairs does not correspond to the degree to which it is good. How strongly one should favour an objectively valuable object depends on the 'distance' between oneself and the object/…/, [and] this distance has many dimensions, including modal distance, temporal distance, and 'personal' distance. It is, therefore, all too crude to say that it is always fitting to feel more strongly about a better state of affairs or to be emotionally indifferent between states of affairs of the same value. Note, however, that it does not seem fitting to *judge* a possible suffering as less bad just because it is a remote possibility, a condition of another person, or something past. (Bykvist 2009, p. 16)

Bykvist considers some possible rejoinders to this distance problem. He dismisses them, however. For example, he would reject the idea that the father in our boat should try to exclude any knowledge of his own personal relations to the children. Bykvist argues that equipping the person who should do the favouring with a veil of ignorance is not in every case a viable response. Such a veil, he thinks, would not work, because in some cases we cannot help having knowledge that would prevent us from disregarding our relations to the evaluated state of affairs; in other words, we would not be able to favour from 'zero distance' (Bykvist 2009, p. 20).

Suppose Bykvist is right about it being somehow impossible not to have this sort of knowledge. I am not sure whether this overturns the veil of ignorance proposal. Perhaps I misunderstand Bykvist, but as far as I understand the veil of ignorance, all that it requires of us is that we should not permit our knowledge to play any relevant role when judging what it is fitting to favour. I am not sure that this is impossible. However, I will not pursue this line of thought here. Instead, I want to say something about the distance problem that relates to what we talked about earlier. Bykvist makes it clear that the FA analysis he wishes to confront with the distance

[16] Besides the challenge to the FA analysis that I consider here, Bykvist's article contains other serious challenges meriting full discussion rather than the brief examination I have been able to provide here.

[17] The distance problem was pointed out by Blanshard (1961, p. 287). I thank Noah Lemos (personal communication) for pointing this out to me.

[18] The Problem, as Bykvist also points out, is in effect a 'wrong kind of reason' problem for the analysis.

problem only acknowledges good, period (Bykvist 2009, p. 16). It might nonetheless be of interest in this context to outline how an enriched version (à la FAP) of the FA analysis accommodating at least two varieties of goodness deals with this sort of case.

Suppose you are, not the father, but a bystander. It seems psychologically credible that at one point or another you might deliberate about saving *A* for his father's sake—imagine, say that *A* is the son of an old friend, now in the boat, and that this is the only thing you know about the children. You might end up thinking that this is what you should do. But, then again, despite realizing that you have some reason to save *A* for your friend's sake, you might reach the conclusion that, morally, you should disregard the fact that you know that *A* is your old friend's child. Perhaps you reflect that you should toss a coin, because you understand that saving *A* has the same value as saving *B*. However, that may be, the idea that you might want to do something for the father's sake reflects, I think, that there is, if you like, more than one value involved in such as case. So just as Bykvist might be right about the common-sense intuition that it is fitting for the father to favour saving *A* more intensely than saving *B,* I suspect that common sense would eventually express this intuition as follows: There is an impersonal good (similar to the kind of goodness we detected in the Spassky–Fisher case), and then there is what is good for the father—and, most certainly, what is good for the children.

An FA analysis embracing good, period, as well as good-for appears to be better equipped to deal with the distance problem. When it is viewed in terms of what is impersonally good and personally good, the boating example ceases to appear paradoxical. It is just a case, among many others, in which we face the option of realizing one kind of value at the cost of another kind of value.

One might object to this value-dualistic approach by pointing out that there is something counterintuitive about the idea that the father should have a reason to favour the rescuing of his child *for his own* sake. It might be fitting for a witness to save *A* for the father's sake. But, surely, it is not becoming of a father to have this sort of attitude: he should save his son for *his son's* sake. We have therefore removed a paradox at considerable cost to our intuitive sense of the father's moral engagement with the situation.

I think there is something in this objection. It would be unfitting for the father to *actually* favour saving his child for his own sake. This would certainly be the case if his sole motivating reason was the idea that the act was done for his own sake; in such a case it would be quite natural to be alarmed. The question is whether having a reason to save his son for his son's sake rules out the thought that the father also has a normative reason to favour saving the child for his own sake? I do not think so. The explanation for this is, I think, mainly twofold. First, in a situation like the one we are imagining here there will be several reasons to act rather than just one. Having a reason to save *A* for his own sake is quite consistent with the fact that he has other reasons to save *A,* including that he should do it for *A's* sake. Also, it is at least plausible to think that these reasons outweigh his reason for *actually* favouring saving the child for his own sake. So although you have reason for a certain kind of favouring, the weight of another reason may well mean that you should in fact not

act on your first, pro tanto reason. So although there is something in the objection, it does not identify a genuinely counterintuitive implication.[19]

At first sight, FA analysts would therefore be well advised to incorporate the notion of a final good-for. The distinction between two varieties of goodness and value allows us to handle the distance problem. There is, as have already said, a price, though. Matters become more complicated once we allow personal as well as impersonal values to play a role in our deliberation. How should we weigh these values against each other? Is it possible to do so, or should we accept that in cases where both kinds of value are involved the situation calls for a choice that will in some fundamental way define who we are as persons? We might ask: Are you (am I) someone who chooses the personally valuable over the impersonally valuable, or do we have different priorities?

These are difficult questions, and for the present, I am not sure what to say about them. However, I am convinced that once Kevin has read these thoughts of mine, it will be fitting for me to take a mulligan and correct my infelicities—for, I venture to say, my own as well as Kevin's sake.

Acknowledgment I am indebted to, Roberta Colonna Dahlman, Noah Lemos, Anne Meylan, Hélène Pessah-Rasmussen, Carlo Proietti, Anne Reboul, Paul Robinson, and Wlodek Rabinowicz for beneficial discussions. Financial support from the Swedish Research Council is gratefully acknowledged.

References

Blandshard B (1961) Reasons and goodness. George Allen & Unwin Ltd, London

Bykvist K (2009) No good fit: why the fitting attitude analysis of value fails. Mind 118(469):1–30

Hobbes T (1969/1889) The elements of law: natural and politic. Frank Cass, Plymouth

Moore GE (1993/1903) Principia ethica. Cambridge University Press, Cambridge

Mulligan K (2009a) Emotions and values. In: Goldie P (ed) Oxford companion to the philosophy of emotions. Oxford University Press, Oxford, pp 475–500

Mulligan K (2009b) On being struck by value—exclamations, motivations and vocations. In: Merker B (ed) Leben mit Gefühlen: Emotionen, Werthe und ihre Kritik. Mentis, Paderborn, pp 141–161

Mulligan K (2009c) Selbstliebe, Sympathie, Egoismus. In: Mulligan K, Westerhoff A (eds) Robert Musil—Ironie, Satire und falsche Gefühle. Mentis, Paderborn, pp 55–73

Prichard HA (1928) Duty and interest (first published as an inaugural lecture). In: Prichard HA (2002), Moral writings, (Reprinted, MacAdam J (ed)). Oxford University Press, Oxford, pp 21–49

[19] In Rabinowicz and Rønnow-Rasmusen (2004), 'The Strike of the Demon: On Fitting Pro-Attitudes and Value', we suggested that in the case of favouring something for the right reasons, reasons might in fact have a dual-role. We favour the object on account of some of its properties. They appear in the intentional content of the pro-attitude. At the same time, they are supposed to make the object valuable. Consequently, they also provide reasons for favouring the object. However, the example of the father apparently suggests that there are cases in which the right reasons should not in fact be part of (or, more cautiously, exhaust) the intentional content of the pro-attitude, even if they are value-making properties.

Rabinowicz W, Rønnow-Rasmussen T (2000) A distinction in value: intrinsic and for its own sake. P Aristotelian Soc 100(1):33–51
Rabinowicz W, Rønnow-Rasmussen T (2004) The strike of the demon: on fitting pro-attitudes and value. Ethics 114:391–423
Rønnow-Rasmussen T (2011) Personal value. Oxford University Press, Oxford
Rønnow-Rasmussen T (2013) Good and good for. In: LaFollette H (ed) International encyclopedia of ethics. Blackwell, Oxford
Scanlon TM (1998) What we owe to each other. Harvard University Press, Cambridge
Williams B (1981) Practical necessity. In: Williams B (ed) Moral luck: philosophical papers 1973–1980. Cambridge University Press, Cambridge, pp 124–131

Chapter 5
Knowledge, Emotion, Value and Inner Normativity: KEVIN Probes Collective Persons

Anita Konzelmann Ziv

> *"The whole doctrine of personalism […] would be ultimately a matter of indifference to ethics if it did not indirectly foster the axiological prejudice […] that the higher values attach to the persons of the higher order […] but to man only the lowest moral values."*
>
> N. Hartmann

Abstract Kevin Mulligan has argued that intuitionism about values is a powerful tool to explain, among other things, "the distinction between what I ought to do and what I must do (practical necessity)" (Mulligan, Leben mit Gefühlen: Emotionen, Werte und ihre Kritik, pp 141–164, 2009). The distinction concerns the difference between moral norms, conceived of as external reasons of acting, and personal norms, conceived of as internal reasons. The kind of intuition the argument relies on is affective and characterized in terms of "being struck by value". One crucial assumption is that affectivity subsumes epistemic states (non-reactive knowledge) and motivational states (reactive emotions). Value feeling is presented as a kind of non-propositional knowledge that can and often does acquaint us with what we value most, with the inner norms or "vocations" that constitute the person we are. The aim of the present paper is to explore to what extent this specific view on personhood that links the knowledge-emotion-value relation of affectivity (KEV) to a personal property of inner normativity (IN) can modify or improve theories of so-called social persons or plural persons. In a first step, I will outline the criteria established for "plural persons" by their advocates. On the basis of these criteria, I will then discuss some reasons for the claim that "plural persons" do have inner norms of the kind mentioned before. In a third step, I intend to show how the KEVIN account interferes with some of the criteria for "plural persons", mainly because of its emphasis on affective knowledge. I conclude that accepting KEVIN either leads to abandoning the claim that plural persons have inner norms or requires the criteria for plural persons to be modified.

Keywords Collective person · Social person · Intimate person · Personal ethos · Vocation · Axiological attitudes

A. Konzelmann Ziv (✉)
University of Geneva, Geneva, Switzerland
e-mail: anita.konzelmann@unige.ch

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_5,

5.1 Introduction

An important claim in social ontology has it that some social units are person-like and ought to be considered as persons. Explicit claims for the personhood of social units have been made by philosophers of the early phenomenological tradition, for example Max Scheler, whose attempts to found an ethical personalism include a theory of collective persons (Scheler 1973a, pp. 519–572). More recently, some philosophers in the analytical tradition have offered accounts of groups in terms of personal subjects. In his "Groups with Minds of their own", Philip Pettit argues that "rational unification is a project for which persons must be taken to assume responsibility" and that, consequently, "social integrates" capable of making avowals of intentional states and acknowledging them as their own "are institutional persons, not just institutional subjects or agents", "on a par with individual human beings" (Pettit 2003, pp. 185, 188). The well-known "plural subject" account of Margaret Gilbert takes a similar vein, suggesting that acts of "joint commitment" to being or doing F "as a body" generate *sui generis* plural subjects of intentional attitudes, states, or actions F. While Gilbert does not strictly qualify plural subjects as persons, she holds that the general idea of a plural subject "goes beyond the idea of a plural subject of goal acceptance or […] acting together" and that the first-person pronoun "we"—in a non-distributive reading—is the standard form of a plural subject referring to itself (Gilbert 2006, p. 166). This at least suggests close vicinity to a view of the plural subject as exemplifying personhood.[1]

Both Gilbert's and Pettit's accounts of personal plural subjects draw on certain essential features of persons. In particular, the capacity to exemplify a variety of different modes of intentionality and the rational unification of the exemplified attitudes, states, and acts, arguing that they apply to groups as well. Both accounts hold, in addition, that a collective person P's state or behaviour F is discontinuous or only contingently continuous with the state or behaviour F of any or all of P's members.[2] In Gilbert's account, continuity is ruled out by the normative force of the act of joint commitment from which the plural subject of F emerges, while Pettit understands

[1] Gilbert sometimes says that the social unit generated in a joint commitment might properly be called a "person": "Quite generally, if Anne and Ben are jointly committed, they are jointly committed to doing something as a body, or if you like, as a single unit or "*person*". Doing something as a body, in the relevant sense, is … a matter of "all acting in such a way to constitute a body that does it". Doing is here construed very broadly. People may be jointly committed to accepting (and pursuing) a certain goal as a body. They may be jointly committed in believing that such and such as a body (Gilbert 1999, p. 147, *my emphasis*). On the other hand, Gilbert uses the qualification "personal" to distinguish individual members' attitudes, states and acts from those of the "plural subject", which seems to suggest that plural subjects are not persons.

[2] The discontinuity claim suggests that all predicates *F* applied to groups are per se "collective", no matter whether they are semantically collective such as *playing a symphony* or semantically distributive such as *going for a walk*. It suggests that application of a concept to a group implies the concept's inevitably falling in the scope of the operator "*cum*" or "*together*", and that this *cum*- or *together*-"modality" inhibits any distributive reference of the concept.

continuity as a constitutive impossibility inhering in the structure of judgment aggregation in terms of which he construes collective rationality.

Whereas the capacity of exemplifying a variety of different types of intentionality as well as the unification of exemplified attitudes, states and acts are widely acknowledged features of personhood, axiological personalism particularly emphasizes the fact that persons are first and foremost axiological beings.[3] Persons carry specific (dis)values, persons have insight into (dis)values, persons are attached to (dis)values and persons realize (dis)values (Scheler 1973a; Hartmann 2007). This is why persons, in contrast to non-personal entities, are "beings who are interested in others" and as such interwoven into a texture of relations of "disposition, conduct and evaluation". The specific personal "attitude" is manifested in "acts of taking sides for or against" each other, e.g. acts of mutually recognizing, bearing ill-will or loving one another. It is precisely this genuine capacity of axiological attitudes and behaviour that distinguishes persons from mere rational subjects and rational agents (Hartmann 2007, pp. 321–324).

On the axiological account, the core or essence of a person is her "individual value-essence" or "ethos", i.e. a specific pattern of values the person is particularly attached to (Scheler 1973a, p. 489). The content of individual value-essence can be experienced as "pointing to 'me'", thus placing the person "in a *unique* position in the moral cosmos" and "calling" on her in a determinate way. Experiencing one's individual value-essence grounds the normative experience of an "individual ought", i.e. an experience "of the ought-to-be of a content, an action, a deed, or a project through *me,* and in certain cases *only* through me" (Scheler 1973a). Scheler comments that this "fundamental experience" of an inner particular normativity is the basis of "the ideas of 'calling' ('vocation'), 'mission', and 'election' for a task" (op.cit. 490, note 121), and that individual vocation obtains independently of whether "the man in whom it is embodied" falls short of responding to its call (Scheler 1954, p 123). The axiological theory of the person is intimately related to a theory of affectivity. Affective attitudes and states are essential for a person's having an individual value-essence, for her knowing values in general and her ethos in particular, and for her being motivated in realizing values. Attachment to values is considered as a basic affective attitude, the lack of which disqualifies an *x* for being a person.

In the following, I shall use the acronym KEVIN for the axiological conception of persons. The letters "I N" stand for *Inner Normativity*, i.e. the function of a person's individual value-essence or ethos to motivate the attitudes and doings of its bearer. The letters "K", "E", "V" stand for the terms *Knowledge, Emotion, Values* and summarize the general idea that knowledge of values is affective in kind as well as the more specific idea that knowing one's ethos is feeling the values one is particularly attached to. Knowability of ethos is the condition of apprehending its normative call. Using the acronym KEVIN for the axiological conception of the person pays tribute to Kevin Mulligan, the person honoured in this volume. As is

[3] In the following, I focus on the axiological personalism developed by Max Scheler. For an overview of different strands of personalism see Bengtsson 2006.

well-known, Kevin Mulligan contributed in many ways to reconstructing and developing the KEVIN account of the person that was outlined by Max Scheler in his *Material Ethics of Values* and taken up by Nicolai Hartmann in his *Ethics* (Scheler 1973a [1913–1916]; Hartmann 2007, 2009, 2004 [1926]). A glance at *A bibliography of Kevin Mulligan's Work* (in this volume) reveals a variety of aspects under which Kevin investigates the panoply of topics involved in the KEVIN conception. The aim of my investigation here is to examine the relation between KEVIN and collective persons. To what extent does the axiological conception apply to collectives? Do groups have and experience an "individual value-essence" as well as an "individual ought" which calls them to do this or that? Can "joint commitment" account for a collective's attachment to values? Can the model of "judgment aggregation" explain the attitudinal property of a collective or group ethos?

5.2 Ethos and Vocation

The colloquial practice of referring to an "ethos" of collective entities such as nations, companies or trade unions suggests a positive answer to the question of whether institutional groups have an individual value-essence which qualifies them as persons in the axiological sense. Consider, for example, the view that identifies a group ethos with the group's "constitutive goals and values, norms, standards, beliefs, practices", which are collectively "endorsed" and ground "group reasons" (Tuomela 2007, pp. 18, 3). This view ascribes a role to group ethos that corresponds to the function of a person's individual value-essence to ground her individual "ought-to-be" and thus to determine her course of action. It seems, however, that in spite of this analogy, attributing an ethos to groups is not sufficient to consider them as persons. Raimo Tuomela's notions of "ethos" and first-personal "we-attitudes" are neither linked with the notion of a "collective person" nor with the notion of attitudes that are discontinuous with the attitudes of individual persons. On the contrary, Tuomela defends a "membership account" of collective intentionality that explains group attitudes in terms of their members' we-attitudes and specific membership relations. This example suggests that group ethos is very well conceivable as a property of which the individual value-essences of singular member persons are constitutive.

Partisans of axiological conceptions of the person likewise credit collectives with an ethos. Scheler reports how the ideas of an "individual ethos of a people and a nation" and of a "peculiar 'national conscience'" were introduced by Schleiermacher, assisted by Herder and Leibniz (Scheler 1973a, p. 513, note 155). Moreover, he indiscriminately attributes a "system of concrete value-assessments and value-preferences" to subjects as different as "an individual, a historical era, a family, a people, a nation, or any other socio-historical group", and refers to axiological systems of this kind as to "the ethos of any such subject" (Scheler 1973b, p. 98). Just like "individual value-essence", "ethos" is intimately related to its bearer's "innermost essence" as well as to affective attitudes: "The fundamental root of this

ethos is, first, the order of love and hate" (Scheler 1973b). If "ethos" is intended as synonymous with the "inner value-essence" that determines an *x* as a person and if ethos is exemplified by collectives, then collectives exemplifying an ethos need to be considered as persons in virtue of their ethos.

Scheler explicitly adopts this view when he outlines an account of collective persons in his "*Formalism*" (Scheler 1973a). It is, however, a theory that conceives of collective persons as bearers of attitudes that are essentially continuous with those of individual persons. On the one hand, this is due to the fact that *person* is defined as a twin-entity consisting of an "intimate" and a "social" person, which are equally fundamental. In virtue of their social twin—their "social person"—individual persons unite into social units some of which display the properties of persons (e.g. nations). Accordingly, the intimate twin of a collective person—its "intimate person"—is the assembly of the social twins of the collective's constitutive members. Individual persons' social twins are no less constitutive of their particular personality than their intimate twins. Therefore, if their social twins unite to constitute the intimate twin of a collective person, properties relevant to their personality are necessarily contained in the collective's properties. On the other hand, the KEVIN conception requires continuity between a collective person and individual member persons because of the constitutive role that affective attitudes play for the ethos of the person. It is very doubtful that affective and emotional attitudes, the core of a person's ethos, entirely result from decisions. Individuals can join their wills to collectively support actions, goals and decisions that they would not perform or defend as individuals. But they can hardly join their wills to *feel* in a way that is contrary to their individual emotions. Since they cannot be implemented by committal acts, the collective affective attitudes required by the claim of a collective ethos seem to call for an account that embraces the attitudes of the individuals involved.

Even if KEVIN is a conception that can allow for collective persons, it is by no means obvious that it requires them. Nicolai Hartmann, an admirer of Scheler's personalist value ethics of which he adopts and develops large parts in his *Ethics* (1926), strongly criticizes the claim that collectives ought to be considered as persons. Acknowledging that "social units in a certain sense are also fulfillers of acts, and that to a certain extent the carriership of ethical fulfilment inheres in them", Hartmann nevertheless doubts "whether this fact alone is sufficient ground for attributing to them personality in the full and intensified sense" (Hartmann 2007, p. 335).[4] The reason for his worry is that a collective's executing tasks, quarrelling or having debts seems to always depend on the initiative of single persons, that communal ends seem to be envisaged by individuals and that wrongdoing and guilt seem to "fall conspicuously upon them" (op. cit 336). The worry, in other terms, concerns the question whether a collective has sufficient ontological autonomy to

[4] In his writings, Hartmann uses the word "*Personalität*" for the property of being a person, which has been translated as "personality" in the English translation of the *Ethics*. Since "personality" is commonly used to refer to one's individual person or character (in German: "*Persönlichkeit*"), I will use this term only in the latter sense and refer to the general property of being a person by the term "personhood". Quotations from Hartmann's *Ethics*, however, contain the term "personality" wherever the translator chose to use it.

count as a person in her own right. And this question, in turn, emerges from the belief that the properties relevant to collective personhood are essentially continuous with the properties relevant to individual personhood, a fact that is considered to set "very definite limits […] to the possible extension of personality" (Hartmann 2007).

Hartmann's reluctance to attribute genuine personhood to collectives apparently derives from his view that the properties determining personhood are principally the attitudes making up one's individual ethos. While he agrees that attitudes relevant to an ethos can aggregate to produce a collective or shared ethos, he considers attitude aggregation as a process whose result is never detached from or discontinuous with the attitudes of the individuals involved. Aggregated attitudes and acts may very well "work like a collective act of a communal person and […] possess value and disvalue", but this is not tantamount to their "centralization in a corporate personality". Rather, the phenomenon of aggregated attitudes amounts to "common participation in the ethos and the ontological and ethical connection among the individual personal subjects", and awareness of this "common possession" of ethos "subsists exclusively in the individuals, and not in the community" (op. cit. 338). The suggestion, then, is that sharing a common ethos or participating in a common ethos immediately follows from aggregating attitudes contained in the ethos of individual sharers. In other terms, ethos sharing is procedural, like the aggregating of attitudes from which a common ethos continually emerges. The primarily affective nature of the attitudes relevant to an ethos, as well as the procedural nature of sharing an ethos, gives a certain plasticity to the common ethos. If Hartmann hesitates to infer collective personhood from the existence of a collective ethos, this is apparently because he considers a collective ethos as insufficiently stable to constitute an autonomous person, or perhaps simply because he thinks that a collective ethos conceived of in terms of "shared attitudes" is a property that does not necessitate a bearer over and above the persons who bear the shared attitudes.

From what has been outlined so far, it follows that attribution of an ethos to collectives does not require us to consider them as persons. Hartmann is right to reject collective personhood on the basis of a distributive view of collective ethos. And this view, in turn, seems to be explicable in terms of the nature of ethos constitutive attitudes. Since they are not necessarily propositional, axiological attitudes such as feeling values and value preferences are not aggregated on the model of rational aggregation of beliefs and desires. They need not, therefore, exhibit the discontinuity between collective and individual stance revealed in applications of this model (List and Pettit 2011, pp. 42–58). Moreover, it is doubtful whether ethos relevant attitudes are suitable targets of "joint commitment", i.e. whether persons can jointly commit to feeling value *V* as a body.

According to KEVIN, a person's ethos determines his life in both non-normative and normative way. Non-normative determination is "mute"; it runs by way of tendencies the person simply exemplifies. Normative determination, however, "appeals to" or "calls on" the person, making her understand what she must do or avoid. The appropriate response to the call of one's ethos is to be motivated to behave in a way that realizes the values revealed in the call. The term "vocation" denotes the specifically normative dimension of ethos, its "voice" by which it expresses

what values "ought to be" for this particular person. Vocations exercise "valuational pulls", which, in contrast to the pull exerted by role obligations and moral laws, are not experienced as external to the person, but "as implicated in the individual's own sense of personal values" (Blum 1994, p. 105). Experiencing the normative power of vocation has the quality of discovering personal values that are not the "product of any self-determination" (Mulligan 2009, p. 148).

KEVIN is an account that not only ties the notion of "ethos" to that of the person, but also the notion of "vocation" to that of "ethos". The axiological perspective, then, seems to require attributing vocations to any *x* to which an ethos is attributed, hence to any *x* considered as a potential person. Attributing vocations to collectives is, however, not as common a linguistic practice as attributing an ethos. We rarely say now of nations, states or cultures that they have a vocation, except perhaps in the case of peoples considered to be chosen by God, or missionary communities. This fact might be explicable by a specifically anthropological feature of vocation made salient by Husserl in "*Erneuerung als individualethisches Problem*" (1924).[5] There, Husserl distinguishes "pre-ethical" from "ethical" vocation; the former regulating a specific domain of one's life (e.g. professional life), the latter one's entire life. Both pre-ethical and ethical vocations are considered to determine "specifically human forms of life" for which the ability to "*survey one's entire life as a unit and to universally valuate it* with regard to realities and possibilities" is constitutive (op. cit. 27, *my emphasis*).

Husserl emphasizes both the epistemic and the normative dimension of vocation when he characterizes it as "a sentiment (*Gesinnung*) of unconditional devotion to valued goals, emerging from their being unconditionally desired" (Husserl 1989, p. 29). On the one hand, a vocation makes the subject discover that he unconditionally desires certain value-goals, and on the other hand, it urges him to devote his life to the realization of these values. Husserl relates both these qualifications to the specifically human awareness of mortality and the limits this impose on human projects. From this perspective, non-human beings seem to be excluded from having vocations, because gods, angels and collectives are not mortal, let alone aware of their mortality. Neither the Windsor family nor the Vienna Philharmonics, the Swiss government, the French State or the Palestinian people can survey the whole of their "life" as a finite whole in a way that calls for devoting it to a value that is not only "appreciated and esteemed", but "wholeheartedly loved from the innermost centre of one's personality" (Husserl 1989). Awareness of transience and real, foreseeable end of existence is not part of plural subjects' worldview. They lack the sense of urgency this awareness confers to what one can achieve in life. If anybody can be literally acting *sub specie aeternis,* it is rather collectives than individual persons. Husserl's account at least suggests that vocation's "call" is unconditional not in the modal sense of impossibility of alternatives (see Williams 1981; Mulligan 2009, pp. 146–151), but rather in the sense of urgency imposed by the human condition. Vocation's call is unconditional because it presents the person to herself simultane-

[5] Husserl 1989, pp. 20–41.

ously as the only one to realize a particular pattern of values and as the one whose life is irrecoverably running out.[6]

Understanding the unconditional nature of vocation in terms of a sense of urgency grounded in awareness of one's existential transience might indeed explain why vocation is not easily attributed to collectives. Alternatively, we might explain this fact by simply holding that the term "vocation" is out of fashion and has been replaced by the term "conscience". Attributing a conscience to groups or collectives is a rather well-established practice that is often related to their being attributed an ethos and personhood.[7] Like vocation, conscience is typically conceptualized in terms of a "voice" whose appeal is "heard", and it is typically considered a "*private* monitor" in the sense that its verdicts are limited to "judgments about the rightness or wrongness of the acts only of the owner of that conscience" (Ryle 1940, p. 31). Conscience is a self-evaluative device that arguably is not simply the mouthpiece of general moral norms or laws, but assesses its owner's intentions and behaviour on the basis of his own particular moral code. On this view, conscience exhibits the same feature of absolute particularity as vocation does, since it "represents the *individual form of the economization* of moral insight" that is "directed to the *good as such* 'for me'", and, consequently, "is *essentially* irreplaceable by any possible 'norm', 'moral law', etc." (Scheler 1973a, p. 324). Moreover, conscience usually speaks or calls loudly when it denounces wrongful behaviour, whereas it is "quiet" when the behaviour monitored is right: "When we say 'Conscience is aroused', we understand immediately that it is set against a certain action. […] [I]t 'warns' and 'forbids' more than it recommends or commands" (op. cit., 322). In this respect, it is similar to vocations which typically reveal themselves negatively: this or that way of life is *not* for me (Mulligan 2005, 2009, p. 148).[8] The important difference, however, is that conscience relates to moral values, whereas vocation is not so limited. A person married for many years may discover by vocation that his mate is not the right partner for him even if he is an attentive husband whose conscience need not accuse him of any wrong behaviour.

In spite of being more frequently applied to groups than vocation, conscience also resists simple collectivization. This is partly due to its self-evaluative nature. If George's conscience calls on him for having cheated in his tax declaration, both the accusing and the responding experiences of "pangs", "stabs" or "twinges" are

[6] Hartmann expresses similar views on the essential relationship between vocation and the human condition when he characterizes man as the "appointed mediator" ("*der berufene Mittler*") between the "realm of reality" and the "realm of the values": Only man has the "clairaudience" ("*Hellhörigkeit*") needed to "discern" the "calling" of values, and only man has the ability to realize their "demands". This "*Weltberuf*" of human beings implies, on the one hand, their having the absolute "freedom of intention attached to ethos", and, on the other hand, their being in the bonds of an "ethos of participation and attending to values" that is "akin to the ethos of love" (Hartmann 1949, pp. 159–174).

[7] Scheler considers the ideas of an "individual ethos of a people and a nation" and the idea of a "peculiar 'national conscience'" as equivalent (Scheler 1973a, p. 513, note 155).

[8] This is compatible with the fact that ethos reveals her "missions" to a person and that mission disclosing is often positive (e.g. "You/I must design functionally perfect buildings!").

his. Compare this to the case of the Christian church praised by the former British Prime Minister Brown as being the "conscience of our country". In what sense is the voice of the church, say in the form of verdicts against abortion or participation in a war, the self-assessing voice of the nation? And how must we conceive of the British nation's experiences of "pangs", "stabs" or "twinges" responding to this voice? Whereas the first question concerns the matter of the legitimacy of an institutional moral authority, the second question concerns the problem of how to account for experiences of institutional bodies. The answers to both these questions must invoke the nation's attachment to a particular set of values, because the concept of conscience requires that *x*'s attachment to a particular set of moral values *V* makes both the call of *x*'s conscience authoritative and its manifestation felt in a specific way by *x*. This suggests that the church is authorized to act as the nation's conscience to the extent that the church enforces the moral values contained in the particular set of values the nation is attached to. By the same token, the nation—because of its being attached to the values invoked in the call—will recognize the church's call as legitimate warning against a certain way of behaving and experience the "pangs", "stabs" or "twinges" resulting from not complying with it. If this is right, it seems that attributing a conscience to collectives presupposes attributing an ethos to them. And if common or collective ethos resists explication in terms of joint commitment and judgment aggregation, we must expect similar difficulties for explications of collective conscience in terms of either of these models.

One such difficulty consists in determining the entitlement to represent the collective in matters of morals. Since the Christian Church is itself a collective body with a particular moral code, it seems as if the Church, in order to represent the nation's ethos, would have to suspend the commitment to its own ethos. Otherwise, the Church would impose its own ethos on the nation, which is not consistent with the role of conscience; a conscience is "the moral voice" expressing the particular ethos of an autonomous moral agent, but it is not a moral agent on its own. If an entity *y*, be it an individual or a collective body, takes the role of the conscience of a moral agent *x*, then y represents *x* and the latter's moral code. Gandhi has been called "the conscience of all mankind" because his conduct was supposed to stand for the values of mankind; Henry Hazlitt has been called "the economic conscience of our country and of our nation" because he was supposed to stand for American values of a more particular kind. If an individual's ethos is exemplary of the ethos of the collective the individual is member of, it is in virtue of this exemplariness of ethos that the member's conduct can successfully represent the collective's conscience.[9] Representative exemplariness of individual ethos implies, however, that individual ethos conforms to collective ethos, and it is implausible that such conformity be a matter of mere contingency. The fact that individuals can function as a

[9] Collective conscience so conceived is explicable in terms of an aggregate attitude that concedes more weight to the attitudes of some particular individuals such as experts or a dictator. These cases require that constraints on aggregation functions such as anonymity and/or systematicity be relaxed (List and Pettit 2011, pp. 42–58).

collective's conscience strongly suggests that individual axiological attitudes are in fact participative of or continuous with collective axiological attitudes.

5.3 Knowing and Realizing One's Ethos

It is not obvious how KEVIN fits collective persons as long as they are modelled on "joint commitment" or "judgment aggregation" accounts. Properties relevant for axiological personalism—like vocation and conscience—essentially involve valuations (*Werthaltungen*), i.e. ways of being engaged in values. Valuations include basic attachment to values, knowledge of values, and being motivated by values. Elsa, for instance, might be attached to the value of perfect musical harmony, even if she never heard an example of perfect musical harmony or else learned about it. Her attachment to this value makes her liable to be "struck" by it, i.e. to immediately know perfect musical harmony when confronted with one of its exemplifications. Elsa's axiological knowledge can take the form of *feeling* values (when she apprehends particular value qualities), or of *preferring* (when she apprehends relations of height between values). Knowing the value of perfect musical harmony she is attached to, Elsa is liable to be motivated to act in ways that will propagate the value that moves her.

Being attached to values, knowing values and being moved by values arguably are not susceptible to being implemented by wilful decision. People can jointly commit to behaving in certain ways, to upholding a maxim, to defending a proposition or to teaching the importance of a value. But they cannot jointly commit to feeling the rightfulness of an acquittal, to being attached to beauty or to being moved by kindness. To the extent that valuations display rather objectual than propositional intentionality—their objects being values, relations between them and goods—valuations also resist to being collectivized on a model that focuses on the aggregation of the propositional contents of beliefs and desires, such as the account of group intentionality developed by Christian List and Philip Pettit (LP-account) (List and Pettit 2011, pp. 42–58). If attitudes of "group persons" are explained in terms of their propositional content, and if basic valuations do not have propositional content, then group persons are devoid of the capacity of basic valuing. From KEVIN's perspective, however, lack of the capacity of affective engagement in values disqualifies an entity of being a person.

Since an ethos is defined in terms of a specific pattern of values a subject is particularly attached to, it is, in principle, knowable by way of feeling and preferring. Affective knowledge of the values one is attached to, is the condition of vocation, i.e. the ethos' "voice", calling on the person that these values ought-to-be. If the motivational force of vocation succeeds to trigger the person's desire to realize these values, it can lead her to form the appropriate intentions. Arguably, these conative valuations or ways of engaging in values are needed to get affective motivation off the ground, even if KEVIN insists that conative directedness to values is grounded in affective valuations (Scheler 1973a, p. 83). Thus, a person's experiences of felt

value (feeling, preferring) are understood as providing the "pictorial or meaning-component" of her characteristic striving and pursuits. This "content" determines the goal of her conations from which springs the "causality of attraction" immediately experienced in striving. Striving to realize one's ethos, then, involves being attracted or "pulled" by its vocational call the felt experience of which determines the striving's content. Simultaneously, striving is experienced as a "push", as "issuing forth" from an emotional state, their "source or mainspring" (Scheler 1973a, p. 344). Hence, axiological personalism claims there is a two-fold dependency of conations' motivational force on affectivity: her striving requires a person to experience the (epistemic) pulling of feelings and preferences as well as the (promoting) pushing of emotional states.

Emotional states, in this picture, appear as a person's affective reactions to her knowledge of value. By their specific quality, they establish for the person *that* the value felt *ought* or *ought not to be exemplified*. Emotional lucidity about what ought-to-be can be complete in itself, i.e. without requiring that the subject ought to do something in order to realize the value that ought to be exemplified. George's sadness about his friend's Elsa's illness reveals to him that the negative value of suffering from sickness ought not to be exemplified, yet his sadness need not reveal to him that he ought to do something about Elsa's suffering. In contrast, George's guilt about his having been nasty to Emma reveals to him that nastiness ought not to be exemplified, *and* that he ought to do something in order to realize the not-being of nastiness. George's "I am really sorry!" toward Elsa is mainly an expression of his emotion of sadness, while his "I am really sorry!" toward Emma expresses a positive action, motivated by his desire to enhance the not-being of nastiness, which in turn is triggered by his emotion of guilt, which reveals to him that nastiness ought not to be exemplified. Motivation, in particular ethically relevant motivation, primarily resides in the emotional ought-(not)-to-be-exemplified reaction to felt value qualities and relations.

In contrast to the standard way of explaining action motivation in terms of desires and beliefs, the alternative proposal in terms of axiological knowledge and emotional attitudes adopted by KEVIN accommodates the requirement of desire independent reasons for ethically relevant action. In order to help his neighbour get a fair trial, George needs neither to desire that his neighbour get a fair trial nor to believe that he can bring about a fair trial for him. If, in contrast, he does not react with indignation to the injustice of an unfair trial, he will not be motivated in an ethically relevant way to help his neighbour get a fair trial.

The fact that affective attitudes towards values can be held without an instance of striving shows that values are not simply dispositions "to be striven for or against" (Scheler 1973a, p. 36), even though desires and conations in general are attitudes which engage in axiological states of affairs. Yet the problem of the relations between conative and affective attitudes towards values remains a thorny one. "Do we first feel the values for which we strive", or "do we feel them in the striving" or perhaps "after the striving, by reflecting on what is striven for" (op. cit., 35) are questions that need to be borne in mind. Consider two more examples of valuational attitudes, which perhaps point to answers to them. Suppose George understands Erna's

indignation about acts of vandalism on the occasion of sport events because he feels the injustice of hooligan behaviour. George's feeling of the negative value of such behaviour need not be accompanied by the desire that this injustice should not exist. Perhaps he has grown so weary of vandalism that he has lost all interest in its presence or absence. Or perhaps he secretly enjoys hooligan actions as nourishing his sensationalism without experiencing, however, any tendency to hinder or promote vandalism. In contrast, George cannot desire that the injustice of vandalism ought not to exist without *feeling* this injustice. In this case, it seems plausible that affective valuation indeed is prior to conative valuation. This finding is even more compelling in cases of value preferences, which reveal an order of values, and need not lead to strivings for or against one of the related values. Eva might prefer George's beauty over Brad's without her desiring that George's beauty rather than Brad's be realized, say in her husband Tim or in her son Peter, or in any other man she knows.

5.4 Aggregating Valuations

If KEVIN is right, a person is essentially constituted of an ethos consisting of particular valuations and the motivations they yield. From this axiological perspective on personhood, a group or collective, in order to count as a person, needs to be liable to exhibit the affective and conative attitudes which are constitutive of an ethos. On first glance, functionalist conceptions of persons as agents, particularly as performers of speech acts, seem to meet this requirement easily, because they usually adhere to a principle of self-ascription. According to such a principle, speech acts, in addition to their fulfilling specific illocutionary roles, are also self-ascriptions of the speaker's underlying intentional states or attitudes (e.g. *x*'s declaring "*p*" ascribes to *x* the attitude of sincerely believing that *p*). Since companies, governments, parties and other institutional groups are undeniably suitable partners in exchanges of ordering, promising, requesting and agreeing, the principle of self-ascription seems to provide these collective subjects of speech acts at one go with the whole array of attitudes held by natural persons. This is made explicit in accounts of the "performative conception of the person", according to which "to function as a person is to utter words as tokens of one's attitudes" (List and Pettit 2011, p. 172).

A main problem for theories that explain collective personhood on the basis of the performative conception is the disparity in accounting for performing collective speech acts and accounting for collective attitudes. Whereas the former resorts to organizational structures and norms, e.g. authority by proxy, the latter does not seem prone to this kind of explanation. The way a company organizes its procedures of decision-taking and assigns responsibilities and authorization validates the speech act of a designed spokesperson as the company's speech act. But how could these structural features explain a correlative underlying attitude of the company? And, what is more, in what ways could such an organizational attitude be relevant to the performance of the act? List and Pettit's performative conception of plural persons as "Groups with Minds of their Own" is an attempt to explain group

personhood in terms of group attitudes, and to explain the latter as aggregates of individually held representational and motivational attitudes (op. cit. 42). The core idea of the LP-account is that collective decision processes run according to an aggregation function (e.g. majority voting) that maps a distribution of individual beliefs or desires held towards a set of propositions onto the collective attitude held towards these propositions (op. cit. 47–50). Notwithstanding the host of fascinating insights the LP-account provides, there is a reason to doubt that collective decision processes really aggregate individual attitudes such as beliefs and desires. Rather, collective decision processes seem to determine, on the basis of expressions of individual negative or positive attitudes towards a proposition *p*, whether *p* or non-*p* shall have the status of a goal or directive on the collective level. Even if the attitudes of individuals towards *p* may be "regulated" by the collective decision "*p*" and eventually change, this is not necessarily the case. In principle, each and every individual is left with the attitude towards *p* they had before. The collective decision for *p* does not require an additional collective or aggregated attitude towards *p* in order for it to function.

The point can be stressed by considering the extreme case of a one-man decision "*p*" issued by a dictator or hierarchical principal of a group. On the model of an aggregation function that relaxes the anonymity condition (i.e. "all individuals are given equal rights in determining the group attitudes", op. cit. 49), the LP-account accommodates dictatorial one-man decisions as cases of aggregated group beliefs or desires held towards *p*. These cases of a group declaring "*p*" allow that no individual, including the dictator, need to believe or desire p. The obvious absence of an "aggregate attitude" towards *p* in settings like this illustrates the asymmetry between ascribing speech acts and ascribing attitudes to groups. It makes clear that group utterances "*p*", in order to fulfil their function, need not amount to uttering words "as tokens of one's attitudes". Public acts of declaring "*p*" by a representative person or body do all the work needed to implement that *p* is a goal or directive for the collective. A similar point has been made by philosophers of "social acts of the mind", even with regard to the individual instances of the acts that are relevant in group contexts. In their accounts of promising and other social acts, both Thomas Reid and Adolf Reinach untiringly emphasize the fact that what makes a promise create an obligation and a right is not dependent on the attitude taken towards its content (e.g. the intention to satisfy the content), but on the nature of the act itself and the understanding of this nature by the addressee and any other persons involved (Reid 1969, 2002; Reinach 1913). A promise properly given acquires and keeps full normative status in virtue of it being properly made; the promisor is under the obligation to perform the promised behaviour independently of whether his promise is a true or false self-ascription of the intention or desire to do so.

If there is a reason to doubt that propositional attitudes are aggregated in decision processes, there are even more reasons to doubt that non-propositional valuations are aggregated along the lines of the LP-model. The closest we can come to real aggregation of attitudes in LP-account is what they call "group deliberation" that "may transform individual attitudes so as to make them more cohesive" (op. cit. 52). Yet group deliberation, they claim, *precedes* aggregation of attitudes.

Margaret Gilbert's account of plural subjects constituted by way of "jointly committing to being or doing F" faces a similar problem. To be sure, Gilbert's theory does not claim that collective attitudes obtain in virtue of aggregated individual attitudes. Her "plural subject" account is an explicit alternative to explanations of collective attitudes, e.g. collective guilt feelings, in terms of aggregative accounts (Gilbert 2002, p. 139). The act of joint commitment that constitutes a plural subject is a conative act, "a kind of joint willing" (Gilbert 2006, p. 225), the normative power of which is "conditional" on the egalitarian relation of mutual reciprocity between its parties. Collective attitudes are taken to be created or generated by this kind of commissive acts. Suppose, for example, that members of a government jointly commit to apologizing for genocide atrocities perpetrated by their compatriots. By the act of jointly committing to apologizing to the victims, the government becomes the plural subject of apologizing. The one act of apology will be borne by one subject, and the apology will count as the apology of the government (or the people represented) and not as the apology of the proxy who performs the linguistic act. This is a very appealing view of what a plural subject is and can achieve. But the self-ascription thesis seems to lurk here too when Gilbert claims that a collective that apologizes or declares being guilty of having done wrong, does in fact believe that it did wrong and does in fact feel guilt for having done wrong, and what is more, the collective holds these attitudes independently of its members' beliefs and feelings (Gilbert 2002).

Valuations by members of the government of collective wrongdoing of their compatriots are most probably fine-grained and differentiated attitudes towards this fact, in accordance with personal perspectives and personal ethos. The fact that the government decides at a certain point in history to apologize for what has been done might depend to a higher or lesser degree on these valuations, but also (and often much more) on other factors of a more pragmatic and strategic nature. There seems to be no conceptual or practical ground why the collective subject of an apology should have particular attitudes towards the fact for which it apologizes. Even less plausible is the idea that by their jointly committing to apologizing, the government members jointly commit to feeling guilt. It is unlikely that anybody can generate an appropriate feeling by a commitment of doing so. And even granted this possibility, there is no compelling ground to believe that it would enhance in any way an act of national apology.

How does axiological personalism fare with the problem of collective attitudes? How does it account for collective ethos or ethos of groups? The most important part of the answer is that a twin-person view, like the one defended by Scheler, strongly suggests an explanation of collective valuations (and attitudes in general) in terms of aggregating individual attitudes. To that extent, it is closer to the LP-account than to Gilbert's theory of joint commitment. The twin-person view contrasts, however, with the LP-account in that it tends to understand aggregated attitudes as "consensual" properties, i.e. properties in which actual individual attitudes literally "participate". Contrary to the compromise reached in processes of collective decision-making, consensus is reached through convergence of attitudes that presupposes their being continuously transformed, for example, in belief revision (Hartmann et al. 2009, p. 111). It is along these lines, in terms of participation and consensus, that

attitude aggregation must be conceived in the framework of the twin-person view, which claims that a person's "intimate person" is given in her feelings of "*peculiar self-being*", while her "social person" is given in specific experiences of herself as bearer "of some personal membership" relation (Scheler 1973a, p. 561).[10] Membership experiences are possible on the grounds of the essentially participative capacity of "co-feeling". In addition, the mutual relation of intimate and social twin is also claimed to be "experienceable" as such within the person (op. cit. 522).[11]

At first glance, a person's ethos or individual value-essence appears to be identical with the "intimate person" that determines her absolute individuality. But on closer inspection it turns out that ethos cuts across the distinction of an intimate and social person. In the axiological perspective, "person" designates what unifies intentional acts and bears a specific set of values (op. cit. 383, 100). Both intentional acts and "values of the person" exhibit the distinction between "social" and "non-social" (op. cit. 519 ff, 566). Accordingly, the part of the person that is a unity of social acts and values is the social person, whereas the one that is a unity of non-social acts and values is the intimate person. One's social person is the subject-centre of one's acts of promising, ordering, respecting or loving, and the bearer of one's values of dignity, honour, or trustworthiness, while one's intimate person is the subject-centre of one's acts of judging, perceiving and willing, and the bearer of one's values of charm, courage and laziness.[12]

Apprehension of non-personal values plausibly is a matter of acts, functions and feelings of non-social intentionality, while apprehension of personal values is social when it aims at personal "values of the other" (*Fremdwerte*) and non-social when it aims at personal "values of oneself" (*Eigenwerte*). Thus, the intimate twin

[10] The intimate person is incommunicable and non shareable, i.e. absolutely alone, and this "absolute solitude […] expresses an *indestructible* (*unaufhebbare*) essential relation of a negative kind among finite persons" (op. cit. 562). The genuine separateness of persons encompasses the aspects of essential *individuality* on the one hand and of absolute *privacy* on the other hand. "Even in our greatest intimacy" with another person, "we know a priori" of the absolute privacy of her intimate person "both that it necessarily exists and that it must remain absolutely inaccessible to any sort of community of experience. The realization that as finite beings we can never see right into one another's hearts […] is given as an essential feature in all experience of fellow-feeling (not excluding spontaneous love)" (Scheler 1954, p. 66).

[11] Scheler uses alternatively the notions "social person" and "collective person" to designate the non intimate twin of a full person. Since the latter expression is also used to denote personal social units, I suggest to use "*social person*" exclusively to denote the twin aspect of being the unified *center of social acts,* and "*collective person*" to denote *social units* having the status of persons.

[12] Scheler's account of the social person is not consistent. On the one hand, the social person is defined in terms of being the author of social acts, on the other hand we read that "the social person first appears as the *bearer* of a peculiar group of *values*" (Scheler 1973a, p. 566), whereby "values of the person" are not identical with "values of acts" (op. cit. 101). Given that the person "exists solely in the pursuance of his acts" (op. cit. 25), values of acts must, however, be intimately related to values of the person. One such intimate relation is manifest in the fact that the values of the social person "'exact' and require specific acts of recognition, esteem, praise, etc.", to the extent that the degree of violation of honor, for example, "is determined by the absence of the social acts" correlative to honor, and not by the "social consequences" that violations of her honor have for the person, nor "by the degree to which one 'feels' one's honor violated" (op. cit. 566).

of George may ponder over his lack of courage while his social twin appreciates Mary's creativeness. George's individual value-essence or ethos, the pattern of values he is particularly attached to and tends to realize, apparently contains social values as well as personal values not exemplified by him. His ethos, then, is the ethos of the entire individual-*cum*-social person in that it determines George in both his intimate and his social being.

Due to the essence of social acts, i.e. their "intention towards a possible community" and consequent "fulfilment" in a community (op. cit. 519 ff), shared performances of social acts constitute the new individual of a real community. Acts such as George's appreciation of Mary's creativeness and Mary's understanding and emotional responding to this appreciation co-constitute the life-community of their marriage. According to Scheler's personalism, actual social units of certain types are themselves persons, called "collective persons", whereas others are not. One of the criteria given for the personhood of collectives is the nature of the core values that a type of social unit exemplifies. "Society", for example, as opposed to community, exemplifies utility and the agreeable. Since both these values are essential values of non-persons, society cannot be a collective person. Nations and cultures, on the other hand, are types of social units that exemplify spiritual values, i.e. values of the person such as honour, dignity or holiness (op. cit. 519–572). Therefore, nations and cultures are collective persons.

In spite of its bizarreness, Scheler's theory of collective persons underscores that if such entities exist, they must exist as the totality of "various centres of experiencing" co-responsibility and co-feeling (op. cit. 520). These "centres" of experiencing essentially participative attitudes—or "social affections" in the terms of Reid—are the individual member persons of the social unit in question. Collective valuations (and collective attitudes in general) obtain only as aggregates of shared actual attitudes of individuals, who in turn can then be said to participate in the collective ethos constituted by their shared attitudes. In fact, this conception of aggregating attitudes seems close to the conception of consensus developed by Keith Lehrer and Carl Wagner. Their theory emphasizes that consensus needs to accommodate not only all individual assessments of the issue at stake but also all mutual assignments of individual trustworthiness and competence. Accordingly, a crucial element of the consensus theory is to account for the weight of "respect" that the parties aiming at consensus mutually assign to each other (Lehrer 2001; Lehrer and Wagner 1980). The interest of consensus theory to systematically integrate mutual assessments of personal weights or values makes it a promising model of how individual valuations might aggregate into collective ethos.

5.5 Should KEVIN Recognize Collective Persons?

The existence of collective persons is not, it may seem, required by assumptions of the axiological personalism KEVIN stands for. If Hartmann, a convinced KEVINist, is right, then collective personalism is entirely built upon the biased heritage of

a rationalism which is disposed "wherever there is a gradation of advancement of form towards cosmic extent not only to transfer subconsciously the attributes of the lower of the only given grades to the higher and more comprehensive but also to magnify them to a proportionately higher degree". This bias leads the rationalist to make the false assertion that collective units must be "persons of a higher potency" because they have some analogies to individual persons (Hartmann 2007, p. 341). If Hartmann is right, then KEVIN can dispense with collective persons.

In particular, KEVIN should not recognize LP-"count-as-persons" of a "bloodless, bounded and crudely robotic" kind who "are not centres of perception or memory or sentience, or even of degrees of belief and desire" (Pettit 2003, p. 188). In spite of their being "conversable", these emotionless "pachydermic and inflexible" creatures respond and reason in a painstakingly tortuous fashion. They "have only a limited range of rights" that leaves no room for "the right not to be owned by others" (List and Pettit, p. 176, 180 f). We understand why KEVIN cannot recognize these disconcerting beings as persons.

If KEVIN recognizes collective persons, then it will be in virtue of their being capable of valuation and their exemplification of axiological properties that are essential for persons. Collective axiological properties can obtain as the results of valuations and axiological properties of individuals that aggregate in the way of consensual properties.

Will KEVIN dispense with collective persons? Will KEVIN recognize collective persons?

It's entirely up to Kevin.

Acknowledgements I am deeply indebted to Kevin Mulligan whose extensive advice helped improve an earlier version of this paper. Knowing Kevin strongly supports my belief that the ethos of an institution, whatever it might be, is participative, i.e. non-contingently continuous with the individual ethos of its constitutive members. Thank you, Kevin, for your commitment and care! I also wish to express my gratefulness to Natalja Deng and Anne Reboul for helpful comments.

References

Bengtsson JO (2006) The worldview of personalism: origins and early development. Oxford University Press, Oxford

Blum LA (1994) Vocation, friendship and community. In: Blum L (ed) Moral perception and particularity. Cambridge University Press, Cambridge, pp 98–123

Gilbert M (1999) Obligation and joint commitment. Utilitas 11(2):143–163

Gilbert M (2002) Collective guilt and collective guilt feelings. J Ethics 6(2):115–143

Gilbert M (2006) A theory of political obligation: membership, commitment, and the bonds of society. Clarendon, Oxford

Hartmann N (1949/1933) Das Problem des geistigen Seins (2nd ed). De Gruyter, Berlin

Hartmann N (2004/1926) Moral freedom (Ethics, vol. 3). English edition: (trans: Coit S, 1932). Transaction Publishers, New Brunswick

Hartmann N (2007/1926) Moral phenomena (Ethics, vol. 1). English edition: (trans: Coit S, 1932). Transaction Publishers, New Brunswick

Hartmann N (2009/1926) Moral values (Ethics, vol. 2). English edition: (trans: Coit S, 1932). Transaction Publishers, New Brunswick

Hartmann S, Martini C, Sprenger J (2009) Consensual decision-making among epistemic peers. Episteme 6:110–129
Husserl E (1989) Aufsätze und Vorträge 1922–1937 (= Husserliana, vol 27). Nijhoff, Dordrecht
Lehrer K (2001) Individualism, communitarianism and consensus. J Ethics 5:105–120
Lehrer K, Wagner C (1980) Rational consensus in science and society: a philosophical and mathematical study. Reidel, Dordrecht
List C, Pettit P (2011) Group agency: the possibility, design, and status of corporate agents. Oxford University Press, Oxford
Mulligan K (1987) Promising and other social acts: their constituents and structure. In: Mulligan K (ed) Speech act and Sachverhalt. Reinach and the foundations of realist phenomenology. Nijhoff, Dordrecht, pp 29–90
Mulligan K (2005) Selbstliebe, Sympathie und Egoismus. In: Mulligan K, Westerhoff A (eds) Robert Musil: Ironie, Satire, falsche Gefühle. Mentis, Paderborn, pp 55–74
Mulligan K (2008) How (not) to become what you are. Unpublished manuscript
Mulligan K (2009) On being struck by value—exclamations, motivations and vocations. In: Merker B (ed) Leben mit Gefühlen: Emotionen, Werte und ihre Kritik. Mentis, Paderborn, pp 141–161
Pettit P (2003) Groups with minds of their own. In: Schmitt F (ed) Socializing metaphysics. Rowman & Littlefield, New York, pp 167–195
Reid T (1969) Essays on the active powers of the human mind. MIT Press, Cambridge
Reid T (2002) Essays on the intellectual powers of man. Edinburgh University Press, Edinburgh
Reinach A (1913/1989) Die apriorischen Grundlagen des bürgerlichen Rechts. In: Schuhmann K, Smith B (eds) Adolf Reinach: Sämtliche Werke, vol 1. Philosophia Verlag, München, pp 147–189
Ryle G (1940) Conscience and moral convictions. Analysis 7(2):31–39
Scheler M (1973a/1913–1916) Formalism in ethics and non-formal ethics of values: a new attempt toward the foundation of an ethical personalism (ed and trans: Frings MS, Funk RL). Northwestern University Press, Evanston
Scheler M (1973b) Selected philosophical essays (trans: Lachtermann DR). Northwestern University Press, Evanston
Scheler M (1954/1912) The nature of sympathy (trans: Heath P). Routledge & Kegan, London
Tuomela R (2007) The philosophy of sociality: the shared point of view. Oxford University Press, Oxford
Williams B (1981) Practical necessity. In: Williams B (ed) Moral luck. Philosophical papers 1973–1980. Cambridge University Press, Cambridge, pp 124–131

Chapter 6
The Argument of Ethical Naturalism

Bernard Baertschi

Abstract Ethical naturalism, the theory claiming that natural facts and especially facts concerning human nature play a justificatory role in ethics, is not very popular amongst moral philosophers. Especially in countries where Kant's influence is large, the charge of naturalistic fallacy is often made against it. The aim of this chapter is to show that this charge misses the point: Every ethical theory is at a certain level based on pure facts, natural or not, and natural facts concerning human nature are particularly suited for this role. The argument in favour of ethical naturalism relies on a concept of human nature that includes basic desires related to ends we ought to pursue, as Aristotle and the Scholastics already saw long ago.

Keywords Ethical naturalism · Naturalistic fallacy · Fact-value dichotomy · Human flourishing · Virtue

6.1 Introduction: Demarcating the Problem

When someboby, especially if he is a philosopher, mentions Ethical Naturalism (*EN* hereafter), we can be almost sure that he will utter negative critical remarks. And if we press him to explain his concern, we will usually not err if we expect, as an answer, the charge of *naturalistic fallacy* (*NF* hereafter). On the other side, ethical books are full of references to (human) nature; we find them even in the writings of David Hume, the father of the contention, as many commentators have indicated (Pidgen 1993). In this chapter, I will try to show that such a charge is misguided and that, contrary to the appearances, *EN* is not only an acceptable philosophical position, but even a good candidate amongst the available options. Of course, *EN* is not without problems, but for the most part, they are exactly of the same nature as those all moral theories encounter.

Let me expand a little this last remark. When *A* objects *x* to *B, B* can answer directly or not. Amongst indirect replies, one consists in showing that *A* commits *x* too. I will term such a reply a *Tu quoque*'s argument. In this chapter, I will have recourse more than is usual to such arguments; they will not show that *EN* is correct,

B. Baertschi (✉)
Department of philosophy, University of Geneva, Geneva, Switzerland
e-mail: bernard.baertschi@unige.ch

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_6,

because they are not able to do it, but they will show that *EN* is not worse than its opponents. I will be more direct and positive too.

To begin with the charge of *NF,* some clarifications are in order. First, what is *EN*? Like all -ism, it has several meanings; but as I will consider *EN* in this chapter, it consists in this main thesis, pertaining to what Anthony Quinton has named 'the central problem of ethics, that is the discovery of a criterion for the justification of judgments of value' (1966, pp. 136–137):

> *EN*: When you are summoned to justify an action, a judgement of value (concerning an action, a behaviour, an institution or a trait of character) or a moral norm, it is *not* inappropriate to invoke a natural fact (more precisely: a natural fact concerning human nature or condition).

It is a thesis about normative justification: what *reasons* do we have to act and to judge as we do? As I will conduct my analysis on this normative level, I will not enter into the metaphysical aspects of the relation between facts and values. More importantly, normative justification is not psychological justification or motivation. Of course, there are many links between them and normative justification is not without effect on the psychology of decision. For instance, Kant demands that our actions conform to the categorical imperative factually (justification) and intentionally (motivation). But the two are conceptually distinct, as is clear when we hear utilitarians argue that the principle of maximisation of utility justifies our actions, but must not be understood as a motive of action. In this chapter, I will never be concerned by the psychological level of motivation.

On this normative level, *NF* states:

> *NF*: When you are summoned to justify an action, a judgement of value or a moral norm, it is *quite* inappropriate to invoke a natural fact.

The reason for that is that it is not possible to derive norms and values from facts. Take this practical inference (*A*):

1. Every human being desires to be happy.
2. Therefore, society ought to promote the happiness of every human being.

Such an inference is a *fallacy,* because it introduces in 2 a deontic verb (ought to) without any corresponding semantic ingredient in 1: So the deontic character of 2 is without justification, it is like manna falling from heaven. To become a valid inference, we must add to *A* another premise. So we have (*B*):

1a. Every human being desires to be happy.
1b. *Human desires ought to be satisfied*.
2. Therefore, society ought to promote the happiness of every human being.

As I hope is now clear, the struggle of *EN* against the charge levelled by *NF* I am concerned with, has nothing to do with the open question argument or with the double nature of thick moral concepts. What I want to deal with is the question of *the sources of normativity*. Kant asks: What is the source of moral legislation? My naturalist answer would be: some natural fact[1]. Of course, not any natural fact: As

[1] It is the title of a book by Christine Korsgaard: *The Sources of Normativity* (1996). Our topic is the same, but not our conclusions.

morality concerns human attitudes and deeds, it will be facts pertaining to human beings, for short, what the philosophical tradition has named 'human nature'. Moreover, not any human fact will be adequate to do this job; as we will see, appropriate facts will be facts internally related to ends or purposes. But I cannot jump so quickly to these conclusions.

It is easy to see that the reformulation of the above practical inference (*B*) does not settle the problem we are confronted with when we are inquiring about moral justification. If *B* is not a fallacy, it is in need of justification too: Why, can we ask, is there an obligation to satisfy human needs? Of course, the problem is here no more a formal one, it is a substantial one, and it is not without interest to note that, often, the charge of *NF* is a hidden way to dismiss a philosophical thesis and to promote another one (Birnbacher 1990, p. 75). So, let us forget *NF* and ask about 1b the same question we have asked about 2: If the deontic character of 2 is justified by the deontic one of 1b, the deontic character of 1b is in turn without justification, it is like manna falling from heaven. So, what could count as a justification of 1b, i.e. of the deontic character of our moral judgments?

Before proceeding, two remarks are still in order:

1. Moral language has two domains: We speak about *values* and about *norms*. In the following I will not explicitely distinguish these two domains when it is not crucial to my argumentation[2].
2. In the definitions of *EN* and *NF* I have given, the notoriously equivocal concept of 'fact' appears, as it is usual in the discussion of the '*fact*-value dichotomy'. This concept has a general sense, as in the definition Chisholm has put forth: 'Facts are the things that make propositions true; if a proposition is true, it is in virtue of a certain fact' (1976, p. 120)[3]. It is not the sense I am interested in, because it could beg the question if we acknowledge moral facts such that 'Human desires ought to be satisfied' is true; in that case, the 'fact-value dichotomy' evaporates. But the 'fact-value dichotomy' is still 'the *is*-ought question'. So, when I speak of 'fact' in this context, I mean 'what can be stated in a descriptive sentence'; for instance 'events' or 'states of affairs' (Rundle 1979, p. 337).

6.2 Some Ends We Pursue Naturally

What does count as a justification of 1b, i.e. of the deontic character of our moral judgments? Human nature, *EN* says.

In the history of western thought, such a thesis has often been voiced. For sure, it had even been the dominant one before Modernity. So it is not necessary for me to begin from scratch and I will start with a thesis professed by scholastic thinkers,

[2] My position on this subject is that values inhabit the most fundamental moral level and that norms are grounded upon values (Baertschi 2008a).

[3] Later, Chisholm reduces facts to true propositions, but it is of no importance for my purpose (p. 123).

from Aquinas onward. It is the thesis that human beings are driven by four fundamental inclinations and that these natural inclinations are the source of morality in human affairs. These inclinations (i.e. dispositional natural desires) are the following:

1. The desire to live
2. The desire to procreate
3. The desire to know
4. The desire to live with human fellows (Timmons 2002, p. 70)

There are similar desires in nonhuman animals, but as animals are devoid of reason, they do not manifest them in the same way. For example, human procreation and animal reproduction aim at the same goal, but they are lived very differently. This is ancient philosophy, but not quite: Sociobiologists too put a fundamental and not necessarily conscious desire at the foundation of morality: The desire to survive, and accept the thesis that our ethics comes from our human nature, a nature that is inescapably social, as Peter Singer states: 'The principles of ethics come from our own nature as social, reasoning beings' (1983, p. 149). It is simply another manner to state 1 and 2, and even 4, as human beings cannot survive alone, without being members of political communities. The difference between sociobiology and medieval Aristotelism is that they adopt different conceptions of nature (and of laws of nature): For the former, it is the modern and scientific conception of it, while the latter entertains an ontological or metaphysical one. It is very important to notice that metaphysics is about facts and not about norms, because we frequently hear people classifying all that is not scientific as ethical or normative. We shall see that other naturalists resort to the psychological side of human nature, invoking natural sentiments like benevolence. But always, it is human nature.

It is interesting to note that sociobiology too has been charged for committing *NF*. Daniel Dennett replies by a kind of *reductio ad absurdum*: 'If "ought" cannot be derived from "is", just what *can* "ought" be derived from? Is ethics an *entirely* "autonomous" field of enquiry? Does it float, untethered to facts from any other discipline or tradition?' (1995, p. 467) It is the manna's argument, and it is a good argument: ethics must regulate *human* behaviour, so it cannot be completely deconnected of facts about human beings and nature. But in my opinion Dennett stops too early and does not take seriously enough the manna's argument: 'It is one thing to deny that collections of facts about the natural world are *necessary* to ground an ethical conclusion, and quite another to deny that any collection of such facts is *sufficient*'. So understood, the naturalistic thesis evaporates, because, as we have seen, every practical inference contains some facts about the natural world. Even a Kantian will agree with this passage of Dennett. To be a true adept of *EN,* you must claim that facts are *sufficient* to ground ethics, i.e. there exists some basic practical inference that contains *only* facts and nothing else. Why? Is not ontological naturalism, i.e. the thesis that only natural facts exist, a sufficient ground for *EN*? No, because you can be an ontological naturalist and a moral contractualist as you can be a Platonist and a moral naturalist. *EN* is not even internally linked with foundationalism, because you can be a coherentist and accept that amongst the subsets

of beliefs in your maximally coherent set of beliefs, there is one subset containing only propositions about natural facts that is sufficient to justify all the moral beliefs you entertain. No, the reason why facts must be *sufficient* is, as Dennett states clearly, that morality cannot stand alone and makes itself the justifying work. But we must be careful not to propose a '*greedy* reductionism', that is a reductionism which 'rush from facts to values' to quote Dennett once more.

Let us come back to desires. A desire is teleologically structured, it is internally related to an *end*. So the doctrine of fundamental inclinations says that (a) there are ends we pursue naturally and (b) these ends are somehow normative.

I think that the first part of this thesis is uncontroversial. Of course, the pursuit of these ends takes many forms, depending on the social and cultural surroundings we live in; but this is quite natural too, because of 4. So our argument starts, and controversies with it, as soon as we ask, following the second part of the above thesis: Are there natural ends we *ought to* pursue?

6.3 Natural Ends We Ought to Pursue

This too is in a sense not very controversial: Even Kantianism which is usually considered as an anti-teleological doctrine, presents some ends as obligatory, i.e. the perfection of oneself and the happiness of others. And it is not surprising, because action is internally linked with end: every action aims at something. Therefore, the question is: *Which* ends ought we to pursue? Kantianism could nevertheless disagree with the *natural* character of those ends, because it considers nature as essentially related to inclinations and desires that have nonrational aims. Implicitly, this is already an answer to the question: which ends? But we must not be too hasty.

So, to the question 'Are There Natural Ends We Ought to Pursue?', philosophers give frequently an affirmative reply, but differ as to the content of this reply. For a sociobiologist like Michael Ruse, this end is survival, for an egoist like (maybe) Bernard de Mandeville it is personal interest, for David Hume and the Scottish school it is the good of others, through the natural feeling of sympathy; for an eudaimonist like Aristotle it is human flourishing and for a Christian it is holiness. Nevertheless, are those ends really *obligatory*? and if obligatory, are they really *natural*? and if natural, is it because of this character that we *ought to* pursue them?

To pass from matter of fact to duty is to commit crude *NF*. Moreover, it is often a very dubious philosophical move. Think of the position of classical utilitarianism. In an attempt to justify the moral imperative that we ought to maximise the happiness of all, it is sometimes argued that if we must do it, it is because we all want to maximise our own happiness: Normative utilitarianism is justified by universalizing psychological hedonism. But the truth of psychological hedonism is not evident, to say the least.

More deeply, the fact that there are certain ends we ought to pursue does not seem to be unambiguously tied with their natural character. As Elliott Sober states:

'I want to suggest that to the degree that "natural" means anything biologically, it means very little ethically. And, conversely, to the degree that "natural" is understood as a normative concept, it has very little to do with biology.' (1986, p. 234) For a Kantian, this is evident in the radical sense that 'natural' has nothing to do with ethics, as Mats Hansson noted: 'According to Kantian ethics, however, an argument that something is contrary to nature is not an ethically justified reason to prohibit it. Kant […] affirmed, however, that everything perhaps, ought not to have happened which according to the course of nature has happened and according to its empirical ground was inevitable.' (1991, pp. 182–183) But this attitude is not limited to Kantianism. Tristram Engelhardt once said: 'How […] could one hold on nonreligious grounds that homosexuality is *unnatural* if human nature is the product of evolutionary processes that may even have developed genes for homosexuality? […] In any event, the outcomes of evolution would be without intrinsic normative force.' (1986, p. 6, italics mine) But that was his view as a (Kantian) liberal; as an orthodox Christian, his position is rather different: 'Homosexuality, adultery, and fornication may in the context of this world be biologically normal and wholesome in the sense of being adaptive', but nevertheless it does not deprive them of their sinful character, because 'carnal desires other than between husband and wife are *unnatural* in being disordered, as aiming away from salvation.' (2000, p. 247, italics mine) Two occurrences of 'unnatural', and two very different meanings; the first refers to biology, it is a descriptive term, the second to eschatological destiny and divine will, and has a prescriptive function. More precisely, the first occurrence of 'natural' refers to some norm too: In his biological meaning, 'natural' means 'normal from the point of view of adaptation', but this is not a *moral* norm for orthodox Christianism. Could it be a moral norm for other versions of *EN*? As we shall see, an affirmative answer requires that we interpret 'nature' in an essentialist way and contrast it with a third concept, that of 'nature' understood in a completely nonnormative manner, that is nature as the class of everything that we can encounter in physical reality (let us name this concept the *all-encompassing* one).

Acts of will, says a Kantian, versus facts of nature. But is not an act of will a *fact*? When I say: 'I want you to come soon', I utter an order, but this utterance is a psychological episode, so how can it acquire normative authority? Before examining this question, that is before coming to the heart of our problem, let us pause a while on this point: An act of will is a kind of fact. If we examine the justification ethical theories give to normative authority, we soon note that *every* moral doctrine refers ultimately to some fact. Society ought to promote the happiness of every human being. Why? Because every human being *desires* to be happy and human desires ought to be satisfied. But why? Because God *wants* us to love one's neighbour. Or because we *have* freely *decided* to help our fellow human beings when in need. Or because we *feel sympathy* for our fellow human beings when in need. Or because it is in *our interest* to help them (tomorrow, we will perhaps need their help in turn). Or because of our fundamental *inclination* to live in society. Each time, we invoke a justifying fact; even Kant speaks of the categorical imperative as a *fact* of reason (i.e. something tied with our rational *nature*) and his commentators do not hesitate

to speak of rational anthropology[4]. So, in every moral doctrine, practical inferences have the same structure, and a structure that does not express a *greedy* reductionism:

1a. Description of the situation, observations.
1b. Justifying fact.
2. Normative conclusion.

It follows that differences between moral doctrines do not consist in differences between the form of the arguments they employ, but between the content of the justifying premise. They disagree about the nature of the justifying fact: Which fact has normative force or authority?

A very widespread answer is, as we have already seen: *volition,* divine or human. That is not surprising, because in human affairs, will is the power to issue orders and commands. Of course, any act of will does not possess this property; to possess it, will must have acquired authority. But will is, so to speak, the natural bearer of authority.

Another widespread answer is: *reason*. Reason is linked with norms (the norms of rationality, and for many philosophers, rationality and morality are intimately related) and with authority, too. For Kant, will is simply practical reason. Usually, we distinguish the authority of will (deontic authority) from that of reason (epistemic authority) (Bochenski 1979, p. 62), but not infrequently people who possess this second form of authority think that it gives them a deontic one, too (paternalism is a good example of that).

Compared with will and reason, human nature seems a poor candidate. Of course, like Engelhardt and the Scholastics, we can lend it on God, but this move is not allowed to a sociobiologist or to a modern neo-Aristotelian. How then could nature bear normative authority? At first sight, naturalism appears to be hopeless: Everything is in nature, but the task of morality is precisely to make discriminations between what is acceptable or permissible or obligatory and what is not. Let us elaborate on this argument.

In the first chapter of *A Theory of Justice,* John Rawls draws a distinction that has become classical between utilitarianism and deontologism. Utilitarianism, he says, seems more promising, because it proposes a conception of nonmoral good or value (a monistic one: pleasure) and defines moral rightness as the maximisation of nonmoral good: 'The good is defined independently from the right, and then the right is defined as that which maximizes the good' (1971, p. 24). William Frankena makes the same point, generalizing: 'A teleological theory says that the basic or ultimate criterion or standard of what is morally right, wrong, obligatory, etc. is the nonmoral value that is brought into being.' (1973, p. 14) By contrast, deontological theories do not. So, in teleologism, you begin with an axiology (a theory of values, values

[4] See Kant (1788, p. 53): 'Auch ist das moralische Gesetz gleichsam als ein Faktum der reinen Vernunft, dessen wir uns a priori bewußt sind und welches apodiktisch gewiß ist, gegeben'; and Pogge (1998, p. 194): 'Kant takes for granted a general understanding of the laws of (human) nature or of the permanent conditions of human life.' But, as we have seen with Dennett, it is difficult to draw a precise conclusion from that, because we do not know if this general understanding is put in 1a or in 1b. For *EN* it must be 1b.

being properties of what is called 'a good') and then define norms (duties): *values precede norms*. On the contrary, in deontologism, you begin with norms, and what has value is determined by norms: *norms precede values*. But, if classical, this story is not quite correct, because it does not take into account the distinction between nonmoral and moral values. Take for instance the value of pleasure and the value of life. Nobody, be it a deontologist or a teleologist, will deny that we spontaneously or naturally value pleasure or life: pleasure and life are natural or nonmoral goods. Morality begins when we say *which* nonmoral good is to be promoted, protected or honoured (for those predicates, see Pettit 1993, p. 231), that is which nonmoral good acquires the status of a moral one (which nonmoral good counts in or for morality) and so becomes a normative principle of action (Korsgaard 1996, pp. 24–25). Will and reason can do that, but nature? It seems hopeless, because natural belongs to nonmoral: A natural good is a nonmoral or premoral good. So a nonmoral good is a good we often value naturally; we are wired to value it and, as is well known, we are wired to value a lot of dubious moral things.

But, looking more closely, it is not as hopeless as that, because:

1. Will and reason are not such great sources for morality.
2. The different meanings of 'nature' give us some hope.

First, the perennial theological debates about the priority in God of his will over his reason and the Calvinist thesis that God could have ordered us to hate one's neighbour show that will alone may not be a good candidate. And this is stressed if we move to the human realm: Will must neither be arbitrary, irrational, nor evil, for it to have moral authority—Kant distinguishes sharply between *Wille* and *Willkür;* John Harsanyi excludes irrational and evil preferences from the felicific calculus and Christine Korsgaard states: 'The ability to reflect puts the will in a position of self-command' (1996, p. 220. See also Kant 1788, p. 38, and Harsanyi 1977, p. 55). Should we then conclude that rationality is the rightness-conferring property? It is the contention of many philosophers from the Kantian and Utilitarian school (Donagan 1977, p. 215); but this claim too has been hotly contested.

Second, some domains of naturality seem not so badly suited as bearers of moral authority. Think of moral sentiments and especially of sympathy and benevolence, emotions and traits of character that are at the source of morality in David Hume and the Scottish school[5]. The Scholastics we have mentioned spoke of four fundamental inclinations that possess moral authority, and it seems not an absurd idea to give our desire to live a high moral standing; is not this desire in good place to justify the right to life? More generally James Griffin thinks that moral force belongs actually to some desires: 'We have to get behind desires and expectations to the deeper considerations that show which desires and expectations have moral force' (1986, p. 40). 'Behind', because not all desires and expectations are normative, but only those linked with basic or fundamental human needs. What then is the criterion

[5] See for instance Hume (1748, p. 178): 'No qualities are more intitled to the general good-will and approbation of mankind than beneficence and humanity, friendship and gratitude, natural affection and public spirit, or whatever proceeds from a tender sympathy with others, and a generous concern for our kind and species.'

to classify a need as basic or fundamental and why is the basic character of these needs conferring value? Ultimately, the naturalist answer is: because they are tied to our *essential* nature. So, those natural ends we ought to pursue and cultivate would be those ends that are tied to our essential nature.

6.4 Ends that Are Tied to Our Essential Nature

Essentialism is not a well-accepted doctrine; usually it is even rejected without argumentation. But on reflection it is easy to see that nobody can escape a soft form of essentialism, because as Bochenski told long ago in a debate in France with Quine, to separate the essential from the accidental is only to recognize that there exists different strata in (our apprehension of) reality (1962, pp. 184–185). If we, as a species, could not discern the essential from the accidental, we would have disappeared for long! In the same spirit, everybody acknowledges that all the ends we pursue have not the same importance for the person we are (psychologically) or for the person we ought to be (morally).

But what is our essential nature and how can it have moral authority? Alasdair MacIntyre, a well-known naturalist from the Aristotelian camp, states: 'There is fundamental contrast between man-as-he-happens-to-be and man-as-he-could-be-if-he-realized-his-essential-nature. Ethics is the science which is to enable men to understand how they make the transition from the former state to the latter.' (1985, p. 52) The argument is plainly Aristotelian and distinguishes clearly between nature as what there is and nature as what we ought to strive to (our end or telos). And nature-as-telos, it is said, bridges easily the gap between 'is' and 'ought'. MacIntyre, maintain Stephen Mulhall and Adam Swift, 'sees the concept [of telos] as vital to morality understood as a rationally justifiable or objective enterprise, because it alone can license immediate transitions from statements of fact to statements of value or obligation—transitions from "is" to "ought".' (1996, p. 79)

Douglas Rasmussen makes the same point: 'The neo-Aristotelian view of human flourishing […] appeals to human nature in two basic ways: (1) it assumes that human nature is teleological, that is, that human beings have a telos or natural function; and (2) it assumes that this natural function has moral import.' (1999, p. 32) And Rasmussen underlies that it is important not to interpret this telos in the frame of the design theory: 'Teleology has a place in nature not because the universe has a purpose or because God has created and endowed each creature with a purpose. Teleology exists because the nature of living things involves the potential that is irreducible for development to maturity' (1999, p. 35). Physical maturity, but moral maturity, too.

This statement of *EN* is still rather abstract, but it is echoed in many themes voiced by philosophers (and laymen). John Stuart Mill contrasts a dissatisfied Socrates with a satisfied fool (and a dissatisfied human being with a satisfied pig); R. G. Frey opposes the capacities of human beings to those of animals: 'While we share many activities with animaIs, such as eating, sleeping, and reproducing,

no combination of such activities comes anywhere near exhausting the richness of normal adult human life, where love, family, friends, art, music, literature, science, and the further products of reason and reflection add immeasurably to our lives.' (1996, p. 207) Those passages are hierarchical in tone: Human beings are superior to nonhuman animals, but it is not what matters to the thesis we examine now: The important thing is that the activities and capacities mentioned are typical to human beings and that there is little controversy on the fact that such activities and capacities represent what is crucial to be a human being.

It is very important to understand the precise limits of this concept of 'nature' and not to confuse it with the biological and the all-encompassing ones. A lot of objections against *NE* miss this point, as it is already visible in Sidgwick when he says: 'We find no design in nature, if the complex processes of the world known to us through experience are conceived as an aimless though orderly drift of change, [… so] I cannot conceive how it can determine the ends of their action, or be a source of unconditional rules of duty' (1907, p. 81). And, no surprise, he adds: 'Every attempt thus to derive "what outght to be" from "what is" palpably fails'. Nevertheless, 30 pages later we read: 'It seems to me, however, more in accordance with common sense to recognize—as Butler does—that the calm desire for my "good on the whole" is *authoritative*; and therefore carries with it implicitly a rationale dictate to aim at this end' (1907, p. 112). But, as MacIntyre objects to Hume and the Scottish philosophers, it does not suffice to assert the authoritative character of some desire, because it is in need of justification, and from a naturalist point of view, only essential ends linked to human nature can do the work (MacIntyre 1985, p. 49)[6].

EN's argument is now clear: Action has a teleological structure and morally good action must aim at certain ends rather than others. What characterizes actions as right is their contribution to the ends that are essential to a human being in that they are characteristic and crucial to the being he is (Hurka 1993, pp. 9–14). Those ends are therefore normative (in the sense that they are the source of moral authority: *authoritative*), they give our deeds a moral direction and provide justification for our moral beliefs. Truly, it is the life itself of living beings that is teleological: It aims at certain ends and the possession of those ends is not innate (although the capacity to reach them is innate and characteristic of living species). In a certain sense, this Aristotelian argument rejoins the sociobiologist position: Each living being aims at survival, he wants to continue to live, but living the life of the being he is, that is, flourishing—biological nature becomes normative as far as it is tied with essential ends. Of course, ultimate essential ends are not objects of choice, as Aristotle stated (every human being wants *naturally*—so nonvoluntarily—to flourish) and as Anthony Flew (1967, pp. 143–148) critically remarked against some sociobiologists (if survival is our natural end, it is nonsense to urge us to survive), but the means and ways to aim at them—that is intermediate ends—are.

Even a neo-Kantian like Korsgaard presents the normative question along those lines: 'A human being is an animal whose nature is to construct a practical identity

[6] David Wiggins makes a similar point against naturalists who look for the basis of ethics in 'brute nature' (1976, p. 183).

which is normative for her. […] When some way of acting is a threat to her practical identity and reflection reveals that fact, the person finds that she must reject that way of acting' (1996, p. 150). An animal is a teleological entity: its nature imposes tasks on him because it wants to flourish. A human being is an animal of a peculiar sort, endowed with free will, so he can choose his identity (i.e. the ends that are essential for the being he wants to be); but not any identity can do the job and as a *moral* being, he must endorse the identity of a denizen of the Kingdom of Ends.

Now, if you ask: What ends exactly our nature enjoins us to pursue, disagreements will emerge between naturalist philosophers; but if you ask a Christian: What does God enjoin you to do? or a contractualist: What are the duties the social contract make obligatory?—disagreements will be frequent too. My point is only to show that human nature is as good a candidate as God, Will or Reason as the source of moral authority.

Several philosophers disagree. At this point, one argument is often voiced, pretending that human nature is not appropriate to explain the *deontic* character of morality. How can natural facts be at the basis of obligation? We can find such a charge in the writings of Richard Hare and of Charles Larmore. I cannot here answer to it in all details, because it would be necessary to investigate the nature of normative utterances and the relations between values (that are tied to ends) and obligations (Baertschi 2001, pp. 69–86). But I think that the teleological character of human life, that enjoins us to become really human, can be a first good answer (nature gives us a *task*). Another part of the answer will soon be given, when I will speak of the requirements *necessary* to lead a good life.

You may fear that there will be a high price to pay for adopting *EN*: If our nature sets what we ought to do, what we ought to do will be in our interest, because fulfilling his nature is in the best interest of the doer. But is not morality essentially altruistic? I hope that a short answer will be enough here: yes, it is self-interested, but self-interest is an inescapable feature of morality. The Christian hopes to live in paradise for ever if he acts rightly; the contractualist wants to make a contract, that is an agreement that is for the benefit of all, including himself; the utilitarian is not altruistic, but impartial, and, finally, why be altruistic if the interest of human beings were of no importance? (Singer 1979, pp. 208–216). Of course, nothing here implies that self-interest must *motivate* actions; as I have said in introduction, nothing in my argument will pertain to the psychological question of moral motivation.

An action is right if it contributes to the realization of ends that are essential to a human being in that they are characteristic and crucial to the being he is, I have said. But in ethics, a lot of duties and rights we mention seem not to be tied to the essence of *man,* in that they pertain to basic needs we share with many nonhuman animals, like needs concerning life, food, shelter and freedom from pain. Moreover, I have argued that the criterion to tell basic from nonbasic needs was that the first were tied to our essential nature (as human animals); but if we follow MacIntyre, this claim misses the moral point, because what he says pertain much more to *perfectionist* needs than to basic ones. To do justice to this objection, I shall introduce a new distinction that will bring my naturalist's argument to an end.

6.5 The Two Layers of Morality

MacIntyre's realization argument is illustrated by Mulhall and Swift in the following way: 'We can move immediately from the knowledge that a knife is blunt and bent to the conclusion that it is a bad knife, and from the fact that it is sharp and evenly balanced to the judgement that it is a good knife' (1996, p. 79; see MacIntyre 1985, pp. 58–59.)[7]. But it is a little too hasty as a rendering of the argument from function (*ergon* in Aristotelian language). A bad knife *is* a knife, but a picture of a knife is not a knife (remember the commentary on the famous painting of René Magritte: 'This is not a pipe'). If you make a 'knife' with paper (for that you would practice origami), it is not a *bad* knife; it is not a knife at all. To be a knife, an object must have the *function* of a knife: It must be able to cut; if it has not this capacity, it is not a knife, but something else, depending on the function it has. It is exactly the same with human beings: A morally bad human being (a moral fool, to speak *à la* Mill)—an evil one—is nevertheless a human being (and not a pig): He has human telos, but he does not succeed to attain it well. Rational and volitional powers like autonomy are capacities attached to the *possession* of this human telos, and they are at least partly constitutive of the *ontological* status of human beings, often named his *moral* status (Baertschi 2008b, pp. 77–78). Because of the developmental character of human life, human beings, that is beings possessing a human telos and the capacities to attain it, can fail to reach it and stop somewhere in between.

This developmental character implies human functioning at two ends: beginning and achievement. For the beginning, some goods must be available (like food and shelter, but like liberties, too); without them, the capacities to lead one's life cannot be put in action, and this beginning can repeat itself several times (think of health and the institution of health care to help human beings, as far as is possible, to function well and to be 'repaired'). That is the lower layer of morality, so important for human rights and for human dignity (it is the source of many duties towards others, and maybe towards oneself). Once in action, those capacities to lead one's life can achieve their results, that is function, badly or well; whence a second layer (an upper one) of morality. It is the realm of *virtues* and of *perfectionist* values, as Thomas Nagel has called them (1979, p. 554)[8].

In those two layers, normative authority resides in human nature, that is in what Scholastics termed fundamental inclinations (their list or another's, it does not matter). We can observe something analogous even in positions that are neither teleological nor naturalist in their structure. Think of what John Rawls says about primary goods: They are (nonmoral) goods that 'every rational man is presumed to want' (1971, p. 62) whatever plan of life is his own; therefore such goods are

[7] To take a knife as a comparison underwrites the fact that concepts of moral evaluation are of the same kind as concepts of technical evaluation: both are practical and descriptive, as Anthony Quinton showed (1966, pp. 122–123).

[8] See also Mark Timmons (2002, p. 68): 'The goodness of a knife, then, is its being in a state of perfection, and to be in a state of perfection is to be able to perform its function well'. It is a remark on Aquinas' ethics.

goods that every human being needs in order to live a life characteristic of a human being—Griffin asserts explicitly that basic needs 'depend not upon this or that person's particular wish or purpose, but upon something deeper and objective—human nature' (1986, p. 42), even if he does not grant them the same importance as Rawls in ethics. In such a conception, what is left for justice is not the list of goods or their importance for contractants, but the manner to distribute them well when there are not enough of them in a situation of relative scarcity.

Briefly, an entity can function as a human being or not, and if it does, it can function well or not. Functioning is, for a human being, functioning humanly, and good functioning is flourishing; to flourish, human beings must have essential needs satisfied, basic as well as perfectionist ones. As Griffin states: 'What count are what we aim at and what we would not avoid or be indifferent to getting' (1986, p. 22). Moreover, we are *moral* beings, so we cannot flourish as human beings without fulfilling our duties towards ourselves of course (think of gluttony) but towards others, too.

But why towards others too, and not only towards oneself? The naturalist's answer is rather short: because we are so wired that we have concerns for our human fellows. You may express it by invoking natural sympathy, the survival value of cooperation or the fourth fundamental inclination (echoing the Aristotelian saying that a human being is naturally a social being), but those refinements do not change the thesis. If a critic finds such an answer too swift and too short, the naturalist will reply that even those philosophers who have the most carefully tried to justify our duties towards others have not infrequently offered finally a similar justification, if not in general, nevertheless for certain duties. Take first Kantianism. The categorical imperative expresses the requirement of universalizability and is forcefully opposed to inclinations and sensibility that are irreducibly particular and selfish; but when Kant comes to imperfect duties, and especially when he asserts that the moral imperative enjoins us to help our human fellows when in need (the duty of beneficence), his argument is only one of reciprocity (a kind of survival value): It is possible that in the future I need myself some help, so I must help others now[9]. Take then utilitarianism. Its three pillars are maximization, impartiality and sentience. But why do I have the duty to promote the happiness of *all* sentient beings? Mill, after having proposed this inference: 'No reason can be given why the general happiness is desirable, except that each person, so far as he believes it to be attainable, desires his own happiness. This, however, being a fact, we have not only all the proof which the case admits of, but all which it is possible to require, that happiness is a good: that each person's happiness is a good to that person, and the general happiness, therefore, a good to the aggregate of all persons' (1863, p. 36), concedes that

[9] '[…] indem der FäIle sich doch manche eräugnen können, wo er anderer Liebe und Teilnehmung bedarf' (1785, p. 281). Herman (1984, p. 143) has offered a more rationalist explanation of Kant's position on beneficence; but she too acknowledges that it is not deconnected from our condition of vulnerable and imperfect beings: 'It is a fact our nature as rational beings that we cannot guarantee that we shall always be capable of realizing our ends unaided, as it is a fact of our nature that we need things and skills to pursue our ends. […] But we can call on the skills and resources of others to supplement our own.'

it is not possible to give a genuine proof for a first principle and, finally, resorts to 'conscientious feelings of mankind' (1863, p. 30) widespread in humankind. Well, human nature in disguise (love of mankind, that is benevolence) on the basis of a form of naturalism: Happiness is normative because everybody wants to be happy and because we feel sympathy with others. Rawls makes the same point when he states, discussing classical utilitarianism: 'Men's natural capacity for sympathy suitably generalized provides the perspective from which they can reach an understanding on a common conception of justice' (1971, p. 186).

Because of the fragmentation of value, some domain of it may not be covered by the argument from teleological nature, except in a trivial sense: When confronted with any normative demand, we can always answer that we are so wired that we tend to respect it. But this is too short an answer. Think of rights, of respect of human dignity or of deference to nonhuman animals interests. Of course, it is possible to use the Kantian strategy of indirect duties: Looking for personal virtues obliges us to respect rights, dignity and animals, but this does not do justice to those moral demands. If the question of rights is rather easily dealt with, for rights can be justified by basic needs and fulfilling basic needs is a requisite of normal human functioning, it is not the case for human dignity and respect for animals. Those moral demands remind us that primary goods and perfectionist values are not alone in the realm of ultimate intrinsic values and duties towards other beings shows that, ultimately, we human beings are called out by those values. This paves the way for another answer: If we, human beings, react to moral demands as we do (if we are so wired), is it not because there exists in some sense a realm of values that *ought-to-be,* to use an expression of Max Scheler, a realm to which we are naturally sensible? Is this not the reason why Emanuel Levinas gives such an importance to the *face* of others?

6.6 Conclusion

By these last remarks, I do not want to suggest that every moral doctrine is a type of ethical naturalism, openly or in disguise, even if it seems to me that it is not possible to develop a complete moral theory without any recourse to some claim about human nature in the centre of the justification's process (in 1b and not only in 1a). Kant himself, when he contrasts conflictually nature and reason, excludes in fact not so much human nature as such than its nonrational part. But there is a long way from that to a thorough naturalism and its fundamental principle of telos.

In the end, I hope to have made a good case in favour of the thesis that *EN* is not in any worse position than the other main ethical theories, because:

1. Every moral theory needs to call for some basic factual premise in the process of justification.
2. The conception of a teleological nature is not problematic if properly constrained.
3. Duties toward others are not more difficult to account for in naturalism that in other moral theories.

Acknowledgments I am grateful to Yves Page and Raffaele Rodogno for their helpful comments.

References

Baertschi B (2001) La place du normatif en morale. Philosophiques 28(1):69–86
Baertschi B (2008a) Needs and the metaphysics of rights. In: Düwell M, Rehmann-Sutter C, Mieth D (eds) The contingent nature of life. Springer, Berlin, pp 89–96
Baertschi B (2008b) The question of the embryo's moral status. Bioeth Forum 1(2):76–80
Birnbacher D (1990) Rechte des Menschen oder Rechte der Natur? Studia Philos 49:61–80
Bochenski J (1962) Discussion du *Mythe de la signification*. La philos anal Cahiers de Royaumont Philos IV:170–187
Bochenski J (1979) Qu'est-ce que l'autorité? Cerf, Paris
Chisholm R (1976) Person and object. George Allen & Unwin, London
Dennett D (1995) Darwin's dangerous idea. Simon & Schuster, New York
Donagan A (1977) The theory of morality. The University of Chicago Press, Chicago
Engelhardt HT (1986) The foundations of bioethics. Oxford University Press, Oxford
Engelhardt HT (2000) The foundations of Christian bioethics. Swets & Zeitlinger, Lisse
Flew A (1967) From is to ought. In: Caplan A (ed) (1978) The sociobiology debate. Harper & Row, New York, pp 142–162
Frankena W (1973) Ethics. Prentice-Hall, Inc., Englewood Cliffs
Frey RG (1996) Medicine, animal experimentation, and the moral problem of unfortunate humans. Soc Philos Policy 13(2):181–211
Griffin J (1986) Well-being. Clarendon, Oxford
Hansson M (1991) Human dignity and animal well-being. Almqvist & Wiksell, Uppsala
Harsanyi J (1977) Morality and the theory of rational behaviour. In: Sen A, Williams B (eds) (1982) Utilitarianism and beyond. Cambridge University Press, Cambridge, pp 39–62
Herman B (1984) Mutual aid and respect for persons. In: Guyer P (ed) (1998) Kant's groundwork of the metaphysics of morals. Rowman & Littlefield, Lanham, pp 133–164
Hume D (1902/1748) An enquiry concerning the principles of morals. Clarendon, Oxford
Hurka T (1993) Perfectionism. Oxford University Press, Oxford
Kant I (1913/1785) Grunglegung zur Metaphysik der Sitten. In: Cassirer E (ed) Immanuel Kants Werke, vol IV. Bruno Cassirer, Berlin
Kant I (1914/1788) Kritik der praktischen Vernunft. In: Cassirer E (ed) Immanuel Kants Werke, vol V. Bruno Cassirer, Berlin
Korsgaard C (1996) The sources of normativity. Cambridge University Press, Cambridge
MacIntyre A (1985) After virtue. Duckworth, London
Mill JS (1972/1863) Utilitarianism. Dent & Sons, London (ed: Acton HB)
Mulhall S, Swift A (1996) Liberals and communitarians. Blackwell, Oxford
Nagel T (1998/1979) The fragmentation of value. In: Rachels J (ed) Ethical theory. Oxford University Press, Oxford, pp 553–564
Pettit P (1993) Consequentialism. In: Singer P (ed) A companion to ethics. Blackwell, Oxford, pp 230–240
Pidgen C (1993) Naturalism. In: Singer P (ed) A companion to ethics. Blackwell, Oxford, pp 421–431
Pogge T (1998) The categorical imperative. In: Guyer P (ed) Kant's groundwork of the metaphysics of morals. Rowman & Littlefield, Lanham, pp 192–207
Quinton A (1966) Ethics and the theory of evolution. In: Caplan A (ed) (1978) The sociobiology debate. Harper & Row, New York, pp 117–141
Rasmussen D (1999) Human flourishing and the appeal to human nature. Soc Philos Policy 16(1):1–43
Rawls J (1971) A theory of justice. Oxford University Press, Oxford
Rundle B (1979) Grammar in philosophy. Clarendon Press, Oxford

Sidgwick H (1981/1907) The methods of ethics. Hackett PC, Indianapolis
Singer P (1979) Practical ethics. Cambridge University Press, Cambridge
Singer P (1983) The expanding circle. Oxford University Press, Oxford
Sober E (1986) Philosophical problems for environmentalism. In: Elliot R (ed) (1995) Environmental ethics. Oxford University Press, Oxford, pp 226–247
Timmons M (2002) Moral theory. Rowman & Littlefield, Lanham
Wiggins D (1976) Truth, invention, and the meaning of life. In: Rachels J (ed) (1998) Ethical theory. Oxford University Press, Oxford, pp 143–186

Chapter 7
Why We Do Not Perceive Aesthetic Properties

Cain Todd

Abstract This chapter examines whether there are genuine cases of aesthetic perception, and hence whether aesthetic judgements depend on the perception of aesthetic properties. My response will be negative. Specifically, I will argue that although our access to aesthetic 'properties' does appear to resemble perception in certain respects, it differs in two key ways from cases of ordinary everyday perception: (a) in its opacity (i.e. its lacking transparency) and (b) in its partly nonattributive phenomenology.

Keywords Perception · Aesthetic · Properties · Phenomenology · Evaluative

7.1 Introduction

It might seem obvious that we perceive aesthetic value, or more specifically aesthetic properties such as beauty, elegance, ugliness, gracefulness, and so forth.[1] Paradigmatic aesthetic objects such as artworks (including music, literature, film, and dance) and natural objects just *look* or *sound* beautiful, harmonious, sublime, gaudy, or clumsy. Of course, there is some debate about just what counts as an aesthetic property (is 'clumsy' an aesthetic property?) and perhaps some doubt too about whether aesthetic properties can be the object of all the sense modalities. What would it be for something to *feel* or *smell* beautiful or elegant?[2] But these subtleties

[1] Perhaps also we can also sometimes see just that something is aesthetically good, without knowing in virtue of what. Such cases complicate matters and fall beyond the scope of the present discussion, though see the view of Sibley discussed below.

[2] See for example Scruton (2007) for a sceptical view about the capacity of tastes and smells to be of aesthetic value. For the contrary view see Sibley (2001, Chap. 15).

C. Todd (✉)
Lancaster University, Bailrigg, Lancaster LA1 4YW, UK
e-mail: c.todd@lancaster.ac.uk

University of Fribourg, Avenue de l'Europe 20, 1700, Fribourg, Switzerland

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_7,

can be left aside for current purposes. Focussing on putatively obvious cases of aesthetic perception—the beautiful look of a picture, the elegant sound of a melody, the sublime appearance of the mountains—I will be concerned to establish whether our aesthetic judgements really do depend on cases of genuine aesthetic perception. My response will be negative. Specifically, I will argue that although our access to aesthetic 'properties' does appear to resemble perception in certain respects, it differs in two key ways from cases of ordinary everyday perception: (a) in its opacity (i.e. its lacking transparency) and (b) in its partly nonattributive phenomenology.

7.2 Aesthetic Perception and Cognitive Penetration

Common sense, one might think, suggests that we have something like simple perceptual access to aesthetic properties. We just look and see the elegance of the dancer, the ugliness of the face, the daintiness of the vase, and the beauty of the landscape. We hear the beauty of the birdsong and the gracefulness of the melody. What else could such properties be, other than perceptual? Naturally, as with all appeals to supposed common-sense, philosophical difficulties creep in quite quickly, but philosophers too have defended the idea that aesthetic judgements are fundamentally perceptual in nature.

One influential voice here has been that of Frank Sibley, who argued that the application of aesthetic concepts required 'taste', a particular sensitivity to aesthetic qualities that is somewhat rarer than our normal perceptual capacities (2001, Chaps. 1 and 3). Whereas everybody (more or less) possesses the capacity to perceive nonaesthetic features—which are just ordinary sensory-perceptible features—apparently not just everybody has the taste required to detect aesthetic features. This is evident in, and at least partially explains, why there is so much and such fundamental disagreement in matters of aesthetic judgement, and hence why the discernment of aesthetic properties might depend upon individual capacities, and/or require practice and expertise. Sibley's remarks on aesthetic concepts and judgements thus resemble those of David Hume, who held that aesthetic judgement required, amongst other things, a particular kind of delicacy of sentiment, a delicacy which could be improved with practice, and which trained experts could be relied upon to exhibit.

Although Sibley does not talk explicitly of aesthetic *perception*, his view can plausibly be construed in this way. One of the important and peculiar features of aesthetic properties and concepts, according to Sibley, was that although aesthetic properties depend upon or supervene on nonaesthetic properties, no necessary and sufficient conditions can be provided for articulating this relation. As Kant held, aesthetic judgements are not 'rule-governed'. Amongst other things, this might explain why the detection of aesthetic qualities is delicate and requires either expertise or at least a naturally refined sensibility. Alternatively, the explanation might be reversed. In any case, if our access to aesthetic properties is in some form perceptual, it involves a different type of perception to ordinary everyday sensory perception.

The view that aesthetic perception is different from 'normal' nonaesthetic perception, and requires certain additional features that may be subject to the possibility of refinement and expertise has been widely accepted by philosophers of art under the form famously articulated by Walton (2004). He argued that in cases of the aesthetic appreciation of artworks, what we perceive and experience is not anchored solely in basic 'uninterpreted' perceptual properties, for it is contoured and coloured by a range of background factors (henceforth 'subject-relative factors') including education, knowledge, practice, culture, imagination, categorisation, comparison, evaluation, interpretation, intention, and so on. In general, what we aesthetically 'perceive' depends on how this compendium of background knowledge and capacities affects the sensory perceptual experiences of ordinary everyday nonaesthetic properties. More specifically, Walton showed that what aesthetic properties a work of art is perceived to have depends on which of its nonaesthetic properties are *standard, variable,* and *contra-standard* relative to the categories in which it is perceived.

In other words, aesthetic perceptual judgements are exemplars of what has become known as *cognitive penetration*. Put simply, this involves our perceptual states (and hence what we perceive) being affected (penetrated) by higher-order 'cognitive' states, such as beliefs, desires, imaginings, thoughts. More formally, Siegel (2012) has characterised one version of the claim for visual experience as follows:

> **CP**: If visual experience is cognitively penetrable, then it is nomologically possible for two subjects (or for one subject in different counterfactual circumstances, or at different times) to have visual experiences with different contents while seeing the same distal stimuli under the same external conditions, as a result of differences in other cognitive (including affective) states. (204)

Indeed, this kind of penetration seems to be required for the perception of what are called 'higher-order properties', which are any properties over and above the uncontroversial basic properties of visual perception, such as shapes and colours. Standard examples include the property of 'being a pine tree', or moral properties. There is currently some debate about whether we do have access to such higher-order properties in perception, and also whether, indeed, cognitive penetration happens, or is even possible. It is not the point of this chapter to address this debate here. But one particular objection to cognitive penetration does bear on the following discussion. For all that the above formulation says, when the penetrating states influence the content of visual experience, they do so by affecting what parts or aspects of the distal stimuli the subjects *fixate on* or *covertly attend to*. For instance, as Siegel herself points out, the following would count as a case of cognitive penetrability:

> Before and after X learns what pine trees look like, pine trees look different to her, and the visual experiences she has under the same external conditions differ in their content. But this is because gaining pine-tree-expertise makes her fixate on the shapes of the leaves on the trees. If a novice fixated the way the expert did, then her experience would have the same contents. The expertise influences experience content, by influencing fixation points. (205)

So, the objection is that supposed cases of cognitive penetration are really just cases of differences in attention, differences that can be explained by, amongst other things, certain forms of expertise. Such cases, then, would indeed involve differences in the phenomenal character of experience, but this would be due solely to the role of attention in changing the representational content of the experience. Alternatively, one might also argue that such cases then involve a judgement allied to the perception which may affect the phenomenal character of your *overall* experience (judgement + perception) without this amounting to a case of cognitive penetration.

Whatever the force of this objection in nonaesthetic cases, it seems to me that aesthetic experiences resist this kind of analysis. On the one hand, appealing to the attentional capacities and foci of aesthetic experts seems insufficient to explain cases in which the expert tells the nonexpert to examine in as much detail as possible all of the nonaesthetic features of an object on which the aesthetic features supervene, without the nonexpert thereby coming to see the aesthetic features. Attention is insufficient because specific background knowledge, imaginative comparisons, categorical and intentional information might also be required to see, for example, the gracefulness of the line in the Matisse canvas. Moreover, one partial explanation of this would be that, as Sibley and Kant maintained, aesthetic judgements are not rule-governed; that is, the relation of dependence between nonaesthetic and aesthetic features cannot be inferred but has to be perceived. No amount of attention to the nonaesthetic features will be sufficient to guarantee the discernment of aesthetic features.

On the other hand, the objection does not seem phenomenologically accurate, since once you have acquired all of the necessary conditions of expertise it seems that the very character and content of your perceptual experience will be different from what it was before, however, your attention is directed. Just as the pine tree *looks* different to the expert, so the landscape *looks* different once you are able to appreciate its beauty. The nature of your experience has changed, in part because it is now valenced in a way that ordinary sensory-perceptual experiences are not. But this need not result from a change in attentional focus. That is to say, its phenomenal character is valenced, and it has partially evaluative content, in addition to its straightforward nonevaluative perceptual content. Or more precisely, its strictly sensory-perceptual content has taken on an evaluative dimension.

So, I contend that aesthetic experiences often, or even always involve, cases of cognitive penetration, and that they are nonetheless perception-like for all that. Indeed, it is important to note that even if one is an antirealist about aesthetic properties, one still has to explain the perception-like phenomenology of aesthetic experience, and one can still appeal to the cognitive penetration of perception in order to account for it. Yet, however, much aesthetic experience resembles everyday cases of cognitively penetrated higher-order perception, there are two key respects in which it differs from such nonaesthetic everyday perception: opacity and nonattributive phenomenology.

7.3 Evaluative Content and Opacity

Philosophers who defend (strong) intentionalist theories of perception—namely, that the phenomenal character of perceptual experiences is fixed by (and perhaps reducible to) their representational content—often appeal to the so-called *transparency* of perceptual experience, in the sense that they appear to us to have no intrinsic properties that are not representational properties (Tye 2008, 25 ff)[3]. When attending to our visual experiences, it is claimed, all we find are the objects and properties in the world that are thereby represented. The experience is transparent to the world that constitutes its representational content.

Unfortunately, not all aspects of the complex phenomenal character of aesthetic experiences play the right kind of representational role even if we acknowledge that they have some bearing on the nature of the content represented. I am not concerned here to decide the merits of strong intentionalism, but rather to note that aesthetic experiences lack transparency in virtue of the fact that they represent *evaluative content* (i.e. evaluative aesthetic properties) whereas the kind of representational content possessed by perceptual states comprises what we might call value-neutral content. Moreover, insofar as aesthetic 'perceptions' involve the kind of subject-relative factors manifested in their cognitively penetrated nature discussed above, including evaluation and interpretation, and insofar as this may result in equally appropriate but different construals of one and the same set of nonaesthetic features, it cannot be that such experiences involve the straightforward *attribution* of (aesthetic) properties to the world in the way that, say, the perceptual experience of colours and shapes does.

To see this, compare our belief in the response-dependent nature of redness, for example, which does not normally affect the way in which the redness of objects appears to us in perception. It merely appears as straightforwardly a property of objects. When we see the red tomato, its redness and roundness seem to us in our visual experience to belong to the tomato, to be properties of the tomato itself and not properties of our experience. That is, we attribute such properties to the object in our visual experience. The phenomenology of such states is in this way essentially *attributive,* a characteristic that naturally accompanies transparency.

In contrast, I suggest, the evaluative importance attributed to the world in aesthetic experience is lent partly by the subject-relative factors on which it is in part based. We play a role in the construal such that the evaluative 'property' attributed in the content of the aesthetic experience is not simply 'to the world' as though the world were transparently represented as possessing such a quality. Both the relevant background factors and the phenomenal features of aesthetic experience play a role in colouring the way in which the features of the world are represented and hence in determining the *evaluative content,* the way in which the nonevaluative features are

[3] Although strong intentionalism does not strictly speaking require transparency (Byrne 2001), its proponents argue that it is best placed to account for it, so any reasons for doubting the transparency of perceptual experience are at least prima facie reasons to doubt the plausibility of the view. See also Speaks (2010).

(evaluatively) experienced. Indeed, this is why it can be so difficult to prise apart the nonevaluative features from their evaluative seemings, and to list the nonaesthetic features in virtue of which the aesthetic features are applied. Most importantly, the lack of transparency in such experiences is evident to the extent that we are aware of the role that subjective phenomenal character plays in aesthetic evaluative content.

Of course, value-neutral content too is obviously a part of aesthetic representational content since it relies on the perception of nonaesthetic features for its input. Moreover, there may be certain constraints on evaluative content imposed by this nonevaluative perceptual content—not just anything can be aesthetically construed in any way whatsoever. There may be, as a matter of psychological or physical fact, some limits on what can be construed as, say beautiful, by human beings as such, or by some subset of them. There may be more or less 'appropriate', or more or less natural, or conventional ways of aesthetically valuing certain states of affairs. Nonetheless, there may be various incompatible but appropriate ways in which some set of nonaesthetic features can be aesthetically 'perceived'. Let me explain.

Recall Walton's view about the category-relative nature of aesthetic 'perception'. In this light, he argued further that in some cases it is *correct* to perceive a work in certain categories, and *incorrect* to perceive it in others; that is, our judgements of it when we perceive it in the former are likely to be true, and those we make when perceiving it in the latter false.

On this picture, it seems clear that there can be cases where incompatible but equally appropriate aesthetic judgements can be made of the very same object, given the possibility of equally appropriate but different evaluative standards or categorization. The picture that emerges is akin to the idea, championed in particular by Scruton (1974), that aesthetic judgements lend themselves to being understood as 'construals'. Construals are articulated in terms of the familiar phenomenon of aspect-perception or 'seeing-as'. They are essentially experiential states—and hence not reducible to mere beliefs or judgements—involving a way that things appear to the subject, and because they involve a relationship in which one thing is seen in terms of something else, they are interpretive or constructive in a way that mere sense perceptions are not. They are, as Roberts (2003) says in the context of defending a construal theory of emotion, a 'hard-to-specify structure of percept, concept, image, and thought' (77).

In this light it is worth quoting what Roberts has to say about the issue of voluntary control, which sometimes emotions are subject to and sometimes not; which is a question of degree; which involves 'the terms in which a subject "sees" the world including changes in the subject's desires and concerns'; and which involves changing patterns of attention regarding the organization of the features of the relevant object or situation. As such, there may be certain constraints, natural or conventional, on how any given object or state of affairs can be construed emotionally.

> A person at whom I am inclined to be angry may be regarded, quite at will, in various ways: as the scoundrel who did such-and-such to me, as the son of my dear friend so-and-so, as a person who, after all, has had a pretty rough time of it in life, and so forth. If these construals are all in my repertoire, and in addition are not too implausible with respect to the present object [e.g. seeing the young-old woman as an odd-shaped pizza], then the emotions

> that correspond to them, of anger, affection, and pity, are also more or less subject to my will … In some situations an emotion may be so compelling that we are … virtually helpless in the face of it. The therapist of friend, by suggesting and fostering other possibilities of construal, may be able to liberate us from it by contributing to our emotional repertoire. Or she may not. (81)

These observations strike me as right about the emotions, but they can also be applied more generally to the case of aesthetic experience. Given the subject-relative conditions to which aesthetic construals are subject, it is more helpful to think of the conditions grounding aesthetic 'perception' in terms of 'appropriateness' rather than truth.

Scruton makes much of the fact that aesthetic appraisal, and aspect perception more generally, centrally involves an imaginative capacity, and I think that a further feature affecting the opacity of aesthetic experiences is the extent to which they—in particular those involving fiction, and arguably the appreciation of pictorial depiction as well as abstraction—involve and depend upon the imagination.

The subjection of imagination to the will, at least in principle, is a familiar claim and a feature commonly taken to be one of the essential demarcating features of imagination. Is this voluntariness reflected and manifested in the phenomenal character of imagining, of forming images? Such questions are difficult to decide, more difficult than most philosophers generally assume. McGinn (2004), for example, seems ambivalent about the issue. He say: 'This is hard to articulate with any precision, but the lability and fleetingness of images is suggestive of their willed character; their "lightness" goes with the vagaries of volition' (167, n.19). Yet, he also suggests that voluntariness may not be part of the phenomenal character of imagery per se but rather part of the overall awareness that comes with reflection on our first-order mental state. At least, that is how the following claim might be interpreted:

> There is some sense in which the phenomenology of images is affected by their voluntariness: what it is like to have them seems affected by the fact that they are products of will; their causation is somehow imprinted on their phenomenology … So what it is like to have an image incorporates the fact that images are subject to the will, but this character of consciousness does not intrude on the intentional properties of the image. (16–17)

The difficulty here resides in deciding what counts as registering in the phenomenology of a particular state, as opposed to the overall phenomenology of being in that state and perhaps simultaneously other states. For instance, how do we separate out clearly the phenomenal character proper to a visual perception, from the overall phenomenology of the same moment incorporating perhaps various other thoughts and feelings about the objects perceived?[4] I can see no easy answers to such questions, but luckily we do not need to attempt them here, because all I wish to claim is that the role that features such as imagination, interpretation, evaluation, expertise, and so on play in determining the evaluative content of aesthetic experiences is thus very unlike the objective attributive phenomenology that comes with the transparency of perception. Once these factors are so centrally involved, and once it is acknowledged both that they may differ between subjects and that these factors

[4] The difficulty is multiplied when the possibility of cognitive penetration is taken into account.

may allow an indeterminate number of incompatible but appropriate construals, aesthetic 'perceptions' cease to look very much like visual perceptions in their attributive content.

Thus, to the extent that we are more or less aware of the evaluative, subject-relative nature of our aesthetic responses, to that extent the less transparent and hence perception-like our aesthetic experiences will seem to us to be.

7.4 Nonattributive Phenomenology

Where there exists a more or less straightforward causal connection that structures perception, providing its truth conditions and ensuring its transparency, attributive phenomenology and mind-to-world direction of fit, the complex make-up of subjective factors involved in the partly interpretive nature of aesthetic construals undermines any direct analogy here. To see this, we need to turn to the recent flurry of discussion about the role of attention in perception and certain problems that it poses for strong intentionalist views of perception.

Recall that such views deny that experiences have intrinsic phenomenal properties that are not representational properties, holding that all phenomenology supervenes on representational content in the following strong way: necessarily, if two *mental events* differ in phenomenology, then they differ in content (Speaks 2010; Byrne 2001). Some philosophers, however, have given examples which appear to show that attention can make a difference to the perceptual phenomenology of two visual experiences without any difference in representational content, thereby directly undermining the strong intentionalist claim. (MacPherson 2006; Speaks 2010; Block 2010) There may be possible replies open to the intentionalist, but my aim is not to adjudicate this dispute, for I wish merely to borrow some apparent features of the phenomenology of attention that emerge from it in order to cast light on the nature of emotional evaluative content (see Nanay 2010 for one line of response open to the intentionalist; for criticisms of other possible intentionalist replies to such examples, see Speaks 2010; Block 2010; Watzl 2011).

The most convincing of these examples, to my mind, and the most pertinent for our topic, is that provided recently by Block (2010). He examines the case of perceiving two Gabor patches, the one on the left having a 22% contrast, the one on right a 28% contrast, and the 6% difference in contrast easily detectable by simply looking at both, with the focus of attention in the centre between each or evenly distributed over both patches. However, when attention is paid to the 22% patch the contrast is increased to the point where it looks identical to the 28% patch. The point is that focal attention changes certain perceptual qualities, like perceived contrast, such that in the two cases (attention to the left, attention to the centre) there are two different phenomenal experiences of the same item with the same relevant instantiated property, i.e. 22% contrast, yet the experiences are different.

In describing the role of attention in his example, Block convincingly claims that attentively seeing and less attentively seeing the same thing—the 22% patch—are

experiences that differ phenomenally but not in the item seen or in its instantiated properties. Borrowing a phrase from Tyler Burge, he claims that attention lacks the 'phenomenology of objectivity': 'The change invoked by changing attention does not look like a change in the world. There is something phenomenally different between the way the attended 22% patch looks and the way the unattended 28% patch looks, even if they are the same in perceived contrast' (53).

It certainly seems to me that the kind of prominence or salience added to perceptual experience by attention is indeed—and appears to us, at least on reflection, to be—a feature of experience rather than a property of the object perceived. Using a similar example, Watzl (2011) has recently also argued that the prominence that is characteristic of an attended object is, unlike colour, *not* experienced as a property of that object, which it has *independently* of our attending to it. He characterises the nature of attention in the following way: 'consciously attending to something does not just consist in being conscious of a certain way the world appears to be (it has a partially nonattributive phenomenology)' (153).

One consequence that might be drawn from these observations is that attention has its own sui generis partially nonattributive phenomenology. One of the salient marks of this phenomenology, I suggest, is its being, like imagination, subject to some extent to the will (cf. Speaks 2010). Indeed, arguably this voluntariness is a natural concomitant of the partially nonattributive nature of attention's phenomenology since this latter is made apparent partly in virtue of the fact that we ourselves contribute, with voluntary shifts of attention, to the salience, prominence, or determinateness that objects in the world 'take on' when they become the focus of attention. Yet it is important to note that this voluntariness is, and can appear or not to us to be, a matter of degree. Sometimes we can deliberately focus on one feature at the expense of another; we can perhaps attend to two things simultaneously; we can switch and oscillate between objects of attention; but we can also have our attention drawn, accidentally or deliberately, to features we had not noticed.

It is this notion, that attention involves voluntariness and has a 'partially nonattributive phenomenology' that, I suggest, we can appeal to in order to understand and perhaps even explain the phenomenology and intentional content of aesthetic 'perception'.

Aesthetic experiences can be thought of as having a partly nonattributive phenomenology, such as attention. Indeed, it may be that there is more than mere resemblance here and that aesthetic experience borrows this aspect of their phenomenology from the role that attention plays therein. It has, after all, been noted by a number of writers that aesthetic experiences serve to capture and consume attention. It would also explain the way in which aesthetic construals can be subject to some degree to the will, by virtue of attending to different features and configurations of features, and in virtue of the role played by imagination in aesthetic experience more generally. I do not want to commit myself here, however, to any one picture of the exact relation between aesthetic experience and attention and wish merely to point to the salient similarities between them.

Having an aesthetic perception of the beauty of *X* does not just consist in being conscious of a certain nonaesthetic set of features in *X,* it is to construe *X* as

beautiful and this construal, I contend, is partly nonattributive. The object *X* certainly has a range of nonaesthetic features in virtue of which we attribute to it the aesthetic features that we do. It is beautiful because, say, of the way its colours and shapes fit, and its lines are elegant in virtue of the fact that they have this particular shape and curve. But, given the non rule-governed nature of this relationship, the way in which these perceptible properties are aesthetically coloured that constitutes the evaluative content of the experience is not guaranteed purely by the nonaesthetic features of sensory perception. Insofar as this colouring is partly a result of and determined by those subject-relative factors already discussed, the construal depends on us, we are doing some of the colouring of the world ourselves, even if some aesthetic perceptions will appear to us to be more constrained by the nonevaluative features (and other factors) and hence more natural or more appropriate than others.

Aesthetic perceptions, I contend, thus have a twofold content, evaluative and perceptual, and the phenomenology of the former is nonattributive while the latter is attributive. That is what makes the overall phenomenology of aesthetic experience sui generis and *partially* nonattributive; at least, on reflection, for how nonattributive any particular experience will appear to us to be will be subject to a number of variables.[5] One particular variable I have already mentioned is the role that imagination plays in some or all aesthetic experiences. And the imagination also has a partly (or perhaps even wholly) nonattributive phenomenology. When visualising something we are aware (in the standard case) of the fact that this object and its properties are not part of the 'fabric of reality' and hence are not attributed to something independent of our mind; rather, when in this state it is phenomenologically evident to us that imagining lacks what Hopkins (2010) refers to as 'directness':

> **Directness** (Hopkins) = A mental state is direct IFF its phenomenology is as follows: (i) one seems to be related to things that exist independently of one's mental state; (ii) those things seem to be constituents of one's mental state; (iii) the nature of that state seems to be almost entirely constituted by the nature of those things.

If, as I maintain, this lack of directness is imprinted on the phenomenology of imagining, that imagery has an essentially (and not merely partially) nonattributive phenomenology, it is also far from obvious that imagining can be transparent, for when we turn our attention to our images, we find some awareness of this feature.

One possible objection to this line of thought would be to argue that we cannot actually pay attention to this lack of directness, and thus we are not aware of it in an attentive way. That is, one might hold that we cannot pay attention to the nonrepresentational features of images, even if their lack of directness and voluntariness is somehow imprinted on their phenomenology. The relevant phenomenological features are not themselves intentional objects of consciousness; they are not intrinsic features of the experience that can be attended to. Rather, awareness of them is peripheral or cognitive—a bit like in perception.

As with all such claims about what is or is not imprinted on the phenomenology of a given state, it is hard to see a persuasive way of resolving the issue, other

[5] I intend the phenomenology of the experience to be understood very broadly, encompassing cognitive reflection on the nature of our experiences, and on the nature of their relation to the world.

than by further phenomenological claims. However, I think a more plausible view of both imagery and perception is what Kind (2003) calls 'weak transparency', namely the view that it may be difficult but not impossible to attend to the relevant phenomenal features. On such a view, the imagery presented in visualisation would be transparently weaker than the equivalent perceptual states, since the features of voluntariness and nonattributiveness seem like just the kinds of things that mark out the state as one of imagining rather than perceiving and hence are characteristics which can be phenomenologically present and hence which we can attend to.

Of course, to demonstrate this convincingly we would need to be confident of our ability to differentiate the phenomenal characteristic proper to the state from those features proper only to the overall phenomenology and which might not belong to the state. An important part of this task would be to differentiate perceptual from non-perceptual attention, for the latter could arguably be seen as a type of mere cognitive awareness of, say, the nonattributive nature of imagining and not an intrinsic property of the phenomenal character of imagining per se.

Without settling these difficult issues here, however, I just want to claim that (a) the more we are aware of our own subjective contribution to the aesthetic construal of the world, the more opaque our experiences, and (b) the less constrained the construals seem to us to be by the relevant nonevaluative features, the less perception-like our aesthetic experiences will be prone to appear to us to be, and the more the nonattributive nature of their evaluative content will be evident. For these reasons, we should not think of aesthetic judgements as involving some special type of aesthetic *perception* in any but the broadest, and least useful, sense of 'perception'.

Acknowledgments This chapter was completed while on research leave funded by the Swiss National Science Foundation. In addition to them, I would like to thank Kevin Mulligan for his always lucid, illuminating, and provocative comments in many discussions on many topics, including the one presented here. I would also like to thank Anne Reboul for her tireless work on this collection.

References

Block N (2010) Attention and mental paint. Philos Issues 20:23–63
Byrne A (2001) Intentionalism defended. Philos Rev 110:199–240
Hopkins R (2010) Inflected pictorial experience: its treatment and significance. In: Abell C, Batinaki K (eds) Philosophical perspectives on picturing. Oxford University Press, Oxford
Kind A (2003) What's so transparent about transparency. Philos Stud 115:225–244
MacPherson F (2006) Ambiguous figures and the content of experience. Noûs 40:82–117
McGinn C (2004) Mindsight: image, dream, meaning. Harvard University Press, Cambridge
Nanay B (2010) Attention and perceptual content. Analysis 70:263–270
Roberts R (2003) Emotions: an essay in aid of moral psychology. Cambridge University Press, Cambridge
Scruton R (1974) Art and imagination. Methuen, London
Scruton R (2007) The philosophy of wine. In: Smith B (ed) Questions of taste. Oxford University Press, Oxford, pp 1–20

Sibley F (2001) Approach to aesthetics: collected papers on philosophical aesthetics (Benson J, Redfern B, Cox J (eds)). Clarendon, Oxford
Siegel S (2012) Cognitive penetrability and perceptual justification. Noûs 46:201–222
Speaks J (2010) Attention and intentionalism. Philos Q 60:325–342
Tye M (2008) The experience of emotion: an intentionalist theory. Rev Int Philos 243:25–50
Walton K (2004) Categories of art. In: Lamarque P, Olsen SH (eds) Aesthetics and the philosophy of art: the analytic tradition: an anthology. Blackwell, Oxford, pp 142–157
Watzl S (2011) Attention as structuring of the stream of consciousness. In: Mole C, Smithies D, Wu W (eds) Attention: philosophical and psychological essays. Oxford University Press, Oxford, pp 145–173

Chapter 8
Literature, Emotions, and the Possible: Hazlitt and Stendhal

Patrizia Lombardo

Abstract Successful literature offers either directly or indirectly rich descriptions of actions and emotions of human beings. I will first examine in this chapter what can be called the general theory of emotions in Hazlitt and Stendhal. Their approach to affectivity stood in opposition to the most common Romantic sentimentalism at the beginning of the nineteenth century: They believed, in fact, that the heart and the reason were not enemies but on the contrary deeply interconnected. Both writers were convinced that most human activities are motivated by emotions but were utterly suspicious of condescending sentimental attitudes, and condemned Rousseau's complacency in his own feelings. I will then focus on a specific faculty of the mind as understood by Hazlitt in his essays, and by Stendhal in his various writings and his novels: The imagination has the capability of combining the real and the possible, and therefore analyzing by conjecture and thought experiments both the past and the future:

Keywords Sentimentalism · Style · Knowledge value of literature · Essay · Novel

> A possible experience or a possible truth does not equate to real experience or real truth minus the value 'real'; but at least in the opinion of its devotees, it has in it something out-and-out divine, a fiery, soaring quality, a constructive will, a conscious utopianism that does not shrink from reality but treats it, on the contrary, as a mission and an invention. (Musil 1979)

Since the 1970s, literary criticism has been dominated first by formalist and then by cultural studies approaches which display a general attitude of contempt for aesthetic value, science, reason, and the truth. Paradoxically, in the last decades, literary critics have avoided confronting what is fundamental in literature: its intrinsic value as literature, and its ability to represent and express human psychology. The ethical and the aesthetic dimensions of literary forms have been neglected in favor of a somewhat vague sociological vision and a repetitive Freudian–Lacanian interpretation; the immense variety of affective phenomena is reduced to a few complexes (Oedipus, castration) or to an empty linguistic game. Fortunately, for the sake of literature, since the 1980s, some analytical philosophers have been interested in the affective dimension of aesthetic experience and of literature, and have been

P. Lombardo (✉)
Department of French and Center for Affective Sciences CISA,
University of Geneva, Geneva, Switzerland
e-mail: patrizia.lombardo@unige.ch

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_8,

reflecting on the emotions elicited by the arts and on the knowledge value of literature. As evident nowadays, several disciplines are undergoing what can be called *the affective turn,* and a new vision of literary and artistic phenomena, revisiting some questions debated since Aristotle and very important in the eighteenth century, is shattering the structuralist, poststructuralist, deconstructionist, and postmodern spell. Several philosophers and some critics reject the idea that literature belongs to a purely literary sphere under the tyranny of language, understood as noncommunication and shallow rhetoric. Literature has a place in real life, addresses audiences, represents and expresses beliefs, desires, emotions, and values that are crucial for human beings. Literature offers knowledge.

In the past epochs, the concern for the worth of literature was typical, and in the modern time, such concern has been vigorously stated by one of the most important writers of the twentieth century, Robert Musil, whose work, together with that of several Austro-Hungarian philosophers and psychologists of the turn of the nineteenth century has been a constant object of investigation by Kevin Mulligan. A few years ago, I had the opportunity of teaching several graduate seminars with him, often focusing on Musil and Stendhal. Recently, we collaborated in the National Center for Competence in Research (NCCR) in Affective Sciences at the University of Geneva. This chapter owes itself to what I have learned from Kevin Mulligan and from our collaboration.

I will claim that literature provides knowledge. After some introductory considerations on the recent debate on this point, and on some ideas from Musil, I will concentrate on Stendhal and William Hazlitt. These two authors of the first decades of the nineteenth century who have much in common with Musil reflected extensively and deeply on the value of literature. They are important for the study of emotions in general, can contribute to a general theory of emotions, and can answer questions about literature and cognition. I believe that the contemporary debate will gain from the reconstruction of the thought of past authors, who, using a different terminology from the one we use now, put forward a number of fundamental principles for the study of emotions and the cognitive value of literature.

The link between literature and emotions is evident. What else does a novel or a poem or a drama do if not describe, express, suggest, or comment on what happens in the mind of various characters, and the unfolding of their actions and emotions in time? Literature is a good exploration of affective life; hence, when successful, it offers some truths about human psychology.

Some contemporary philosophers believe that literature lacks knowledge value, or only has a very weak one. Peter Lamarque, for example, argues in a recent book that literature cannot provide knowledge "because fictional (or imaginary) situations do not provide real data." He considers it a "mistake to suppose that to be serious or reflective a work must in effect teach something" (Lamarque 2009, p. 253). We cannot measure the bodily reactions, emotions, and sentiments of fictional characters; they cannot be tested. Why, then, should we look for data in a novel?

The second argument by Lamarque against the knowledge value of literature is what can be called the fear of banality. In his opinion, the generalizing truths to be found in a novel or a poem are too simple and plain to offer moral knowledge. In

The Philosophy of Literature, he takes the example of Charles Dickens' *Our Mutual Friend:* "But if we try to extract a worldly truth from all this, a truth independent of fictional particulars, such as 'Money corrupts,' we are back to banality. […] The novel's value resides in the working of the theme, not in the theme's bare propositional content." But why does the "working of a theme" prevent the acquisition of knowledge? Is the working of imagination not continuous with the role of imagination in everyday life, for example in the development of sympathy and the grasp of moral value?

Greg Currie, although he seems to believe that the knowledge value of literature is feeble, points out one connection between fiction and emotions; "fictional narratives of real value ought to have some significant relation to what is true, particularly in the domain of human psychology" (Currie 2010, p. 209).

Those who believe that literature can provide knowledge often insist on the richness of literary examples, or sometimes on the empathic effect of a novel or a poem in the reader's mind. An enthusiastic appreciation of literature was expressed by Martha Nussbaum in 1986. According to her, we can find the best analysis of accidents that affect real life in literary works: "Greek tragedy shows good people being ruined because of things that just happen to them, things they do not control. […] Tragedy also, however, shows something more deeply disturbing; it shows good people doing bad things, things otherwise repugnant to their ethical character and commitments, because of circumstances whose origin does not lie within them" (Nussbaum 2001, p. 25).

Ronald de Sousa developed an argument close to Musil's:"The Art of the Possible in Life and Literature represents a kind of possibility that we might call potentialities. They are not merely tied to the modality of some proposition, but represent actual dispositions or potentialities of a particular person or things" (De Sousa 2005, p. 349. Emphasis added).

Some literary critics have argued for the benefits of empathy; for instance, in her *Empathy and the Novel* (2007), Suzanne Keen thinks that the empathy readers feel in reading novels constitutes a real and easy possibility of increasing their disposition towards altruism.

What could be learned from literature about human emotions, vices, and virtues should not be perceived as having quick and direct results. It does not produce immediate appropriate behavior in human affairs. After all, would anyone seriously assert that the psychology of emotions or the various theories of arousal or appraisal, or the collecting of verifiable data, or the accurate explanation of some ethical concepts by philosophers have an immediate and infallible application in improving our morality?

Let us imagine that we read, for instance, the sad story of Benjamin Constant's *Adolphe,* where the protagonist causes so much pain to his lover and to himself because he is incapable of telling her that he does not love her any longer. Very likely, we will gain little, if anything at all, in terms of empirical knowledge, and we will persist in our indecisions as much as Adolphe does in the novel. Nevertheless, we will definitely be pressed to think about emotions, virtues, and vices. Promoting reflection is an important step on the way to gaining conceptual knowledge.

I insist on the power of literature as a means of reflection. Literature does not simply describe and suggest emotion but also develops our imagination. In his 2002 article "The Wheel of Virtue: Art, Literature, and Moral Knowledge" Noël Carroll rejects the philosophical theses that discard the idea that literature (and art in general) could be an instrument of education and of knowledge. He stresses, on the contrary, the way in which narrative forms of art uses the same type of thought experiments that philosophers use, but in a richer and more complex way, since philosophical thought experiments are often very dry, schematic, and succinct.

Following Carroll, I maintain that literature has a definite cognitive value because it entails an important exercise in imagining the possible, both in terms of other people's life and of our own future. Literary thought experiments provide the same type of mental gymnastics in imagination that is to be found in philosophical reasoning.

It is not necessary to enter the debate about thought experiments which may well go back to the Pre-Socratics (see Rescher 1991)[1]. It is enough here to identify the usefulness of these imaginings. Carroll recalls that the thought experiments used in philosophical explanations often take the form of a narrative while functioning as argument. They are a device that can help to refine, "contemplate, possibly clarify, and even reconfigure our concepts, thereby rendering them newly meaningful." They are useful in making distinctions, "proposing possibility proofs, and assessing claims of conceptual necessity" (Carroll 2002).

If thought experiments are valuable in philosophy, they are crucial in fiction, since they constitute the fictional element itself. The whole of *Robinson Crusoe,* for instance, is a thought experiment testing what happens to a man alone on a desert island (like its modern successor, Michel Tournier's *Vendredi*). It could be said that all novels presenting characters in a given situation are thought experiments. But probably, the many thought experiments displayed within a novel are more important than the overall thought experiment framed by the novel. It is the accumulation and the variety of the branching inferences that call for a continuous adjustment of intertwining hypotheses which is important in the novel. On occasions the reader must react rapidly, on occasions his or her responses require more time. In a novel—or in a good novel—myriads of conjectures blend with descriptions, comments, and digressions. Such is the life of any narrative: it presents episodes and actions (physical and or mental), takes time, and solicits the imagination. Because of its temporal dimension, it often provides more accurate accounts of emotions and their temporal profiles than the dry psychological or the philosophical accounts of schematic emotions at a time.[2] Narratives are crucial in order to understand the difference between an emotional episode and an emotion, since they exhibit what a practical experiment can never do: A chain of emotional events, their motivations, and their aftermaths in a network of beliefs, values, and desires. Short-term and

[1] For an accurate debate on several contemporary approaches to thought experiment, see Engel (2011).

[2] Peter Goldie is very concerned with the emotional dynamics and the fact that our emotional life takes the form of a long narrative (Goldie 2000).

long-term affective phenomena are the domain of narration. The temporal dimension of literature allows for a thick and dynamic account of affective life.

8.1 Musiliana

Robert Musil argues that possibilities belong to reality, and that literature can make that link palpable by fully showing the possible, what could (just) be the case. As stated by Ulrich, the protagonist of *The Man without Qualities,* reality awakens possibilities and art has the capacity of enlarging the sphere of what is possible. In a 1911 article, Musil openly defended the thesis of the value of literature for knowledge, somewhat correcting the famous definition by Aristotle. In the *Poetics,* the difference between history and poetry is described as the distinction between particular facts and general facts, with the consequence that poetry is closer to philosophy since it deals with the general. Musil, however, overcomes the opposition between the particular and the general, identifying their dynamic relation in art:

> To be sure, art represents not conceptually but concretely, *not in generalities but in individual cases within whose complex sound the generalities dimly resonate***;** given the same case, a doctor is interested in the generally valid causal connections, the artist in an individual web of feeling, the scientist in a summary schema of the empirical data. The artist is further concerned with expanding the range of what is inwardly still *possible*, and therefore art's sagacity is not the sagacity of the law, but a different one. […] Where art has value it shows things that few have seen. (Musil 1990, p. 7, emphasis added)

For Musil, then, if valuable art shows "things that few have seen," literature must deal with possible truths—of a type that differs from mathematical truth. Literature has to do with possible worlds, like those hinted at by thought experiments. Discussions of the roles of truth and knowledge should take seriously, I suggest, not only the distinction between the general and the particular but that between contingent and essential truths.

Not unlike Musil, both Hazlitt and Stendhal insist on the value of imagination by means of conjectures about the possible. They did not worry about banality because they understood that literature and art has to vividly express some essential truths about possibilities and necessities.[3] Essential truths, when summarized in a sentence, are often banal, but their presentation in literature is not banal since it should be vivid. What is the "vividness" often claimed by those who talk about what prompts aesthetic emotions?[4] Following the lesson of Musil, Stendhal, and Hazlitt

[3] On this theme, Kevin Mulligan refers to the work of Nicolai Hartmann who distinguished between *Lebenswahrheit* and *Wesenswahrheit,* between truths about life which are true to life and essential truths (*Aesthetik*. Berlin: de Gruyter 1953).

[4] See for example: "This rejection [of human love] of ordinary human passion is nowhere more *vividly* expressed than in the Confessions, where Augustine movingly recalls his own delight in earthly love…" (Nussbaum 1999, p. 61. Emphasis added). See also: "So the question is not really how probable something is but how *vividly* it is imaginable" (De Sousa 2005, p. 351. De Sousa underlines). On the theme of the link between reason and emotions see Reboul (2001).

it can be said that this vividness requires the inseparable connection between form and content in a work of art.

It is obvious that literature has to do with accidental or contingent truths. Novels are filled with descriptions of places, people, and events (historically true or allusive to history or completely invented), and particular actions and episodes. But more importantly, literature also has to do with essential truths, the type of truths that are valid in all possible worlds. Essential truths include mathematical truths but also banal and conceptual or evaluative, e.g., moral truths, such as: Pain is something bad, or justice is more important than a good meal, or every emotion involves the representation of an object. How do the banal essential evaluative truths of literature differ from those dissected by philosophers or bandied around by moralists? Values, as has often been pointed out, depend on non-values, on matters of fact. The color and the shape of an ornament make it beautiful. Certain types of behavior and habits make people foolish or make their actions cruel. The link between different values and what makes people, objects, situations the bearers of different values is at the heart of much literary art. First, literary art depicts the variety of what *makes* what is cruel, foolish, cowardly, unjust possess these properties. Secondly, it depicts the immense variety of what might be called the shades of value. Cruelty, foolishness, injustice, and sublimity come in different kinds, as do colors. The foolishness of Emma is a quite distinctive kind of foolishness, so too is the cowardice of Adolphe. Similarly, the cognitive values exemplified by Ulrich have a quite distinctive flavor. The representation of what makes things valuable and disvaluable and of the ensuing shades of value is the representation of essential truths. When successful, such representation possesses the already mentioned qualities of vividness and concreteness. Thus, whether or not, as Currie suggests, literature fails to provide novel contingent truths about human psychology it provides representations of essential, noncontingent truths. When summarized in capsule form, these are invariably banal—such as all essential, nonmathematical truths. But of course, as we have seen, since successful literature is characterized by the inseparability of form and content its essential truths must be the sort of things which cannot be summarized. If you want to understand the variety of foolishness and of what makes people foolish there is no alternative to reading Flaubert or Musil.[5]

Some writers are chiefly aware of their enterprise and struggle to meet their literary targets. Observing their endeavor in their diaries, correspondence, notes, and theoretical remarks scattered across their fictions or essays, we see that they often have a clear idea that a good novel cannot rely only on accidental truths, an accumulation of details which is at best picturesque. Neither could a good novel pile up essential truths; it would be unbearable, such as a series of commandments or pompous aphorisms. Literature which preaches is bad literature. A good novel is the combination of the two types of truths. The same can be said of a good essay.

[5] Isabelle Pitteloud (University of Geneva) discusses at length about emotions, values, and imagination in the novel in her dissertation on Stendhal, Balzac, Flaubert and their theory of emotions, developing via the close reading of literary examples the argument of Kevin Mulligan's article, "From appropriate emotions to values" (Mulligan 1998).

Musil, Hazlitt, and Stendhal constantly reflected on the literary ideal for which they strove, and believed that style was not an ornament of their ideas but the inseparable blending of form and content. Style was to be understood as the balance between essential and accidental truths in terms of invention and composition, as the whole architecture holding in place the unity of form. This type of style is what strikes their readers as something "vivid."

8.2 Hazlitt and Stendhal

Hazlitt and Stendhal were attached to both the eighteenth-century philosophy of sensibility, and to some endeavors of the Romantic Movement that conquered Europe at the beginning of the nineteenth century. They occasionally met in Florence and Paris. Stendhal admired the *Edinburgh Review* where Hazlitt occasionally published and Hazlitt admired Stendhal's *De l'Amour*. They had a similar understanding of important ethical and aesthetic questions.

First of all, they were convinced that the human mind was a complex and unified system in which both intelligence and affects of various types collaborate. Because of their interest in human affectivity as intimately connected to reason, Hazlitt and Stendhal occupied a similar position among the Romantics. They opposed the over-sentimental or sentimentalist conceptions of their famous contemporaries (Coleridge and Wordsworth in the case of Hazlitt, and Chateaubriand and Madame de Staël in the case of Stendhal), and were both suspicious of the writer who greatly fascinated the whole romantic generation, Jean-Jacques Rousseau.

Secondly, Hazlitt and Stendhal were convinced of the variety of affective phenomena. They thought that there were many different emotions, and not just a handful of fundamental emotions. Consistently with this belief, they realized that emotions can be mixed, self-reflexive and wrong, inappropriate, fake, or sham, and that their effects and combinations needed to be studied. Hazlitt and Stendhal's subtle descriptions of a great variety of affective phenomena are richer than the accounts given by philosophical treatises. Stendhal actually wrote a detailed study of love and Hazlitt, of human disinterestedness but they generally constructed their philosophy and psychology of emotions in scattered reflections, comments, and essayistic writing. They aimed to establish taxonomies of affects, to understand their dynamics, the relation between one emotion and another, as well as between emotions and actions and ideas. Hazlitt refined the form of the essay, and Stendhal gave his best accounts of emotions in his novels.

Hazlitt, who published his numerous essays in journals and reviews, which were then collected in volumes (*The Round Table: A Collection of Essays on Literature, Men, and Manners, 1817; Characters of Shakespeare's Plays, 1817; Table-Talk; or, Original Essays*, 1821–1822), wrote about the most diverse subjects: theatre reviews, politics, fine arts, prose and poetry, commenting the works and ideas of his contemporaries, studying human actions, emotions, and motivations. In his first book, *An Essay on the Principles of Human Action* (1805), he rejected Hobbes'

vision of natural human violence and selfishness, without accepting the vision of natural human sympathy as expressed by Adam Smith in his *Theory of Moral Sentiment*. In Hazlitt's view, human beings are not naturally benevolent, but disinterested. Just as for Hume and Smith, the faculty that can produce this non-selfish feeling is the imagination: "The imagination, by means of which alone I can anticipate future objects, or be interested in them, must carry me out of myself into the feelings of others by one and the same process by which I am thrown forward as it were into my future being, and interested in it" (Hazlitt 1998, p. 3. From now on Selected Writings).

Following Shaftesbury and the tradition of the Scottish Enlightenment, Hazlitt (like Stendhal and Musil) thought of the ethical dimension of life as inseparable from the aesthetic dimension. His essays on Shakespeare's plays constitute a form of literary criticism in which the link between art and life is so fundamental that art becomes, as in Aristotle, a means of education, a true discipline: "Tragedy creates a balance of the affections. It makes us thoughtful spectators in the lists of life. It is the refiner of the species; a discipline of humanity" ("Othello," SW vol. 1, p. 112).

Hazlitt's essays on Shakespeare's plays analyze the mind of various characters, the intentions directing their emotions and actions, while considering at the same time their effect on the mind of the spectator or reader. For example, the examination of Macbeth, of his hesitations, and of "the stings of remorse" that assault his mind is contrasted with the analysis of Lady Macbeth: The remark that she shows "obdurate strength of will and masculine firmness" continues with the scrutiny of connected traits of her character and new developments of these traits. In the tradition of the Aristotelian view that a good fiction should, without striking the eyes, produce in the audience a strong effect of pity and fear (*Poetics,* book II; Chap. I, Sect. 8), Hazlitt observed our affective response to Lady Macbeth, to her will and all her powerful negative emotions—ambition, guilt, self-will, determination, hardness of the heart, lack of natural affections:

> She at once seizes on the opportunity that offers for the accomplishment of all their wished-for greatness, and never flinches from her object till all is over. The magnitude of her resolution almost covers the magnitude of her guilt. She is a great bad woman, whom we hate, but whom we fear more than we hate. She does not excite our loathing and abhorrence like Regan and Goneril. She is only wicked to gain a great end; and is perhaps more distinguished by her commanding presence of mind and inexorable self-will, which do not suffer her to be diverted from a bad purpose, when once formed, by weak and womanly regrets, than by the hardness of her heart or want of natural affections. ("Macbeth", SW vol. 1, p. 100)

Occasionally, in *The Pleasures of Hating* for instance, Hazlitt expanded his Aristotelian vision, in order to investigate complex types of emotions—what we would call today "mixed emotions"—trying to determine the webs in which similar and dissimilar affects can combine:

> Nature seems (the more we look into it) made up of *antipathies*: without something to hate, we should lose the very spring of thought and action. Life would turn to a stagnant pool, were it not ruffled by the *jarring interests*, the *unruly passions* of men. The white streak in our own fortunes is brightened (or just rendered visible) by making all around it as dark as possible; so the rainbow paints its form upon the cloud. Is it *pride*? Is *it envy*? Is it the *force*

> *of contrast*? Is it *weakness* or *malice*? But so it is, that there is a secret affinity, a hankering after, evil in the human mind, and that it takes *a perverse, but a fortunate delight in mischief*, since it is a never-failing source of satisfaction. ("On the Pleasure of Hating", SW vol. 8, p. 118; emphasis added)

Stendhal's entire work—from his art history to his analysis of all the sentiments composing different types of love (*De l'amour. Une description complète et minutieuse de tous les sentiments qui composent la passion nommée amour*, 1822), and his novels and various journals—could be understood as a constant "inquiry" into the emotions and values. Already in his correspondence with his sister Pauline (1804–1805), where he wrote of the typical eighteenth-century question of happiness, he advised her that, in order to come closer to it, it is necessary to be sensitive, to study and know human passions, beliefs, and actions: "*Observons donc; cela ne fait qu'augmenter la sensibilité de notre âme, et sans sensibilité point de bonheur*" (Stendhal 1968, Correspondance I, p. 139, to Pauline, August 20, 1804).

He had no doubt that one can find the best analysis of the "human heart" in:

1. Direct observation of people in the salons
2. Good philosophy
3. Good literature

Stendhal urged Pauline to read Molière, Shakespeare, and Hume, and to study the behavior and the passions of her friends and of herself. Since he was concerned with the way in which what he called "passions" could be classified, he identified long- and short-term passions and "habits of the soul."[6] He recommended that Pauline write out lists of passions and speculate about their possible links, compatibility, and transformations. In his *Histoire de la Peinture en italie* (1817), he often uses the phrase: *les nuances des passions*. These nuances encompass two sets of phenomena. On the one hand, there are the tender emotions that distinguish the sensitive modern human beings from the emotionally simple, harsh, and limited ancient men and women; on the other, there are emotions resulting from the increased expressiveness of the great Renaissance art that he admired. Raphael, Leonardo, and Correggio could infuse their paintings with a subtle gamut of emotions and moods detectable in the expressions and gestures of the human figures, and of the affective qualities of objects. Writing about *The Saint Cecilia* by Raphael, Stendhal finds the broken musical instruments scattered on the floor, visible in the foreground of the

[6] Stendhal explains to his sister: "Il y a des passions, l'amour, la vengeance, la haine, l'orgueil, la vanité, l'amour de la gloire. Il y a des états des passions: la terreur, la crainte, la fureur, le rire, les pleurs, la joie, la tristesse, l'inquiétude.

Je les appelle états de passions, parce que plusieurs passions différentes peuvent nous rendre terrifiés, craignants, furieux, riants, pleurants, etc.

Il y a ensuite les moyens de passion, comme l'hypocrisie. Il y a ensuite les habitudes de l'âme; il y en a de sensibles, il y en a d'utiles. Nous nommons les utiles, vertus; les nuisibles vices. Vertus: justice, clémence, probité, etc. Vices: cruauté, etc. Vertus moins utiles ou qualités: modestie, bienfaisance, bienveillance, sagesse, etc. Vices moins nuisibles ou défauts: fatuité, esprit de contradiction, le menteur, l'impertinence, le mystérieux, la timidité, la distraction, etc " (Stendhal 1968), p. 118 (to Pauline, June 1804).

painting, especially moving. As much as the expression of Cecilia and her companions, they eloquently account for the time just before the sudden and overwhelming rapture that has struck Cecilia. She listens to the celestial music of the angels singing in a cloud high up—her eyes are turned towards the sky: "*L'orgue que tient sainte Cécile, elle l'a laissé tomber avec tant d'abandon, surprise par les célestes concerts, que deux tuyaux se sont détachés.*" Stendhal called the phenomenon of attributing the power of expression to objects: "*l'art de passionner les détails, triomphe des âmes sublimes*" (Stendhal 1929, p. 201, 202). Stendhal's comment on the painting is preceded by a short passage on Shakespeare. He found *l'art de passionner les détails* in some simple and vivid utterances used in *Macbeth*; for example, when Ross remarks the absence of the King in the banquet and Macbeth tells the ghost of Banquo that "The table is full." Here, we might say, Stendhal distinguishes what we have already called nuances of value.

In 1827, when asked about the state of philosophy in Paris by his friend Gian Pietro Vieusseux, he replied that besides logics and metaphysics, there was an important branch studying the knowledge and the explanation of the soul (we would say psychology), and more importantly, yet another branch studying "what happens in the human heart when one feels an emotion" (Stendhal 1967, *Correspondance* II, pp. 131–132, to Gian Pietro Vieusseux, December 1827).

As suggested before, Hazlitt and Stendhal occupy very similar positions within Romanticism. They give similar accounts of various affective phenomena: pleasure, self-love, *amour-propre,* egoism and egocentrism, aesthetic experience, sympathy, self-deception, false feelings, and sham emotions. They both show some essential truths that vanity is a vice, they identify several shades of the disvalue of vanity, and that sympathy and love are indispensable for human society and personal happiness. More importantly, they recognize the value of the "heart," but never, unlike the great majority of Romantics, at the expense of reason. Probably the most striking essential truth whose variety they are concerned to bring out is the goodness of sentiment, and the badness of sentimentality.

8.3 Against Sentimentalism

Because they valued sentiment and the arts as the expression of the human mind and as an education in humanity, both Hazlitt and Stendhal mistrusted any excessive expression of sentiment. They saw in Rousseau the paradigmatic example of sentimentalism in feeling and style. Interestingly, they both associated sentimentalism with sham emotions and with oratorical style; sentimentalism in feelings leads to a self-centered vision of the world, and to complacency, conceit, and cruelty. They are concerned to illustrate the variety of these essential possibilities. Sentimentalist style, they think, is typically characterized by an imbalance between form and content, with a large abundance of words in comparison to the ideas expressed. Rousseau embodied these excesses, and, in spite of the seduction of his writings, Hazlitt and Stendhal were equally committed to denouncing the drawbacks and dangers they represented, and continue to represent.

Stendhal invented a neologism to refer to his endeavor of eradicating the influence of Rousseau. He wrote and underlined in his *Journal* on Mai 23, 1804: "*Dérousseauiser mon jugement en lisant Destutt, Tacite…,*" (Stendhal 1981, p. 152) and, in August 1804, he warned his sister that even if Rousseau is admirable she should remember that "he was always in a bad mood." The more Stendhal acquired his own ideas about life and style (what he called, in his *Souvenirs d'égotisme*: *"la théorie du cœur humain*" and "*la peinture de cœur par la littérature et la musique,*" Stendhal 1982, p. 468), the more he rejected Rousseau. On July 1, 1820, he wrote in his diary: "*Même dans les moments les plus tendres et les plus mélancoliques, comme aujourd'hui […], le tour d'emphase de* La Nouvelle Héloïse *me la rend illisible*" (Stendhal 1982, p. 47). The emphatic mode of the novel which for decades had obtained an overwhelming success, cannot attract the writer who aimed at writing with a few clear ideas and used as an epigraph for his *Le Rouge et le Noir* "*La vérité, l'âpre vérité*." It is doubtless, Stendhal's commitment to the epistemic value of the truth and of its bitterness that recommended him to Nietzsche.

Already when he was very young, he was conscious of the danger of an empty rhetoric of sentiments, and protested against the sham sensibility of those who think of themselves as very sensitive. Sham or artificial (unnatural) emotions are typical of verbose writers, such as Rousseau, Chateaubriand, and Madame de Staël who was a fervent admirer of Rousseau: "*Madame de Staël n'est pas très sensible et elle s'est crue très sensible. Elle a voulu être très sensible […] Ensuite elle a mis là-dessus son exagération*." [7]

Replying to Balzac's enthusiastic article on *La Chartreuse de Parme,* and his minor criticism of its occasional harsh language, Stendhal declared:

> Le beau style de M. de Chateaubriand me sembla ridicule dès 1802. Ce style me semble dire une *quantité de petites faussetés*. Toute ma croyance sur le style est dans ce mot….
> Je lis fort peu; quand je lis pour me faire plaisir, je prends les Mémoires du maréchal Gouvion-Saint-Cyr; c'est là mon Homère. Je lis souvent l'Arioste […] Voici le fond de ma maladie: *le style de J.-J. Rousseau, de M. Villemain, de Mme Sand me semble dire une foule des choses qu'il ne faut pas dire, et souvent beaucoup de faussetés*. Voilà le grand mot. (*Correspondance* III, p. 395, to Balzac, October 16, 1840; emphasis added)

Comparable remarks are put forward by Hazlitt. His essay "On the Character of Rousseau" starts with an attack against Madame de Staël who, he argues, misreads Rousseau, and continues with his criticism of Rousseau's exaggerated sensibility:

> Madame De Stael, in her Letters on the writings and Character of Rousseau, gives it as her opinion, "that the imagination was the first faculty of his mind, and that this faculty even absorbed all the others." And she farther adds, "Rousseau had great strength of reason on abstract questions, or with respect to objects, which have no reality but in the mind." Both these opinions are radically wrong. Neither imagination nor reason can properly be said to have been the original predominant faculties of his mind. ("On the Character of Rousseau", Selected Writings vol. 2, p. 90)

[7] "Si Mme de Staël n'avait pas voulu être plus passionnée que la nature et la première éducation ne l'ont faite, elle aurait fait des chefs-d'œuvre. Elle a voulu sortir de son ton naturel, elle a fait des ouvrages pleins d'excellentes pensées, fruits d'un caractère réfléchissant, et il y manque tout ce qui tient au caractère tendre. Comme cependant elle a voulu faire de la tendresse, elle est tombée dans le galimatias" (Stendhal 1968, p. 214 (to Pauline Beyle, August 20, 1805)).

As for Stendhal, who ended his autobiography *Vie de Henry Brulard* with the remark that an over-detailed account hurts tender emotions,[8] excess of any sort becomes a fault. Consciously or unconsciously, those who believe that emotions can be evaluated as rational or irrational, follow Aristotle's principle of moderation, and are suspicious of any type of overstatement. Therefore, in Hazlitt's opinion, Rousseau's extremely acute sensibility, which is one of the causes of his overwhelming success, ends up being morbid and selfish:

> The strength both of imagination and reason, which he possessed, was borrowed from the *excess* of another faculty; and the weakness and poverty of reason and imagination, which are to be found in his works, may be traced to the same source, namely, that these faculties in him were *artificial*, secondary, and dependent, operating by a power not theirs, but lent to them. The only quality which he possessed in an eminent degree, which alone raised him above ordinary men, and which gave to his writings and opinions an influence greater, perhaps, than has been exerted by any individual in modern times, was *extreme sensibility, or an acute and even morbid feeling of all that related to his own impressions, to the objects and events of his life*. ("On the Character of Rousseau", Selected Writings vol. 2, p. 90; emphasis added)

There are many instances of rapprochements between Hazlitt and Stendhal. Furthermore, much could be said about the use of irony in both writers, and the style for which they strove. The familiar style Hazlitt privileged is analyzed in his essay "On Familiar Style":

> It is not easy to write a familiar style. Many people mistake a familiar for a vulgar style, and suppose that to write without affectation is to write at random. On the contrary, there is nothing that requires more precision, and, if I may so say, purity of expression, than the style I am speaking of. It utterly rejects not only all unmeaning pomp, but all low, cant phrases, and loose, unconnected, slipshod allusions. […] To write a genuine familiar or truly English style, is to write as anyone would speak in common conversation who had a thorough command and choice of words, or who could discourse with ease, force, and perspicuity, setting aside all pedantic and oratorical flourishes. ("On Familiar Style", SW vol. 6, p. 217)

The two writers were, therefore, quite alike in terms of their expectations and ethical–aesthetic ideals. Finally, another striking similarity in their method was that they used and analyzed thought experiments, or conjectures, or the activity of imagination (terms which I use here as synonyms). Recalling the notion dear to Musil, the works of both writers are essays in the most literal sense of "attempts." *Essayismus,* according to Musil, is not just a mode of writing but an investigation which reveals the relations between the possible and the real. The Austrian writer declared in "On the Essay" (1914) that for him "ethics and aesthetics are associated with the word essay," since it is the best way of being precise "in an area where one cannot work precisely." Although the essay does not deal with mathematical truths but with the class of essential truths pertaining to the facts of ethics, values, and feelings, "it investigates and presents evidence" (Musil 1990, p. 28, 49).

[8] "On gâte des sentiments si tendres à les raconter en détail" (Stendhal 1982, p. 959).

Stendhal's way of attempting the possible appears in his use of hypothetical sentences; Hazlitt, as we shall see, elaborates upon the "essay" form, pointing to comparisons as the bridge between the possible and the real.

8.4 "Essayismus" and Comparisons in Hazlitt

Rambling is a term that Hazlitt used often to allude to his own essays. Besides the reference to *The Rambler,* a journal founded by the famous Doctor Johnson around 1750, rambling indicates the attempts typical of the essay. The range of Hazlitt's investigations in the essay form is vast. The analysis of wit and of common sense, of originality, of sham emotions, the critique of Coleridge and Wordsworth, idle speculations about the passing of time, considerations about taste, and comments on paintings and political issues—all these various themes are relentlessly studied. He could spot, through an implacable close reading, the inconsistencies of Joshua Reynold's theories of painting, and "ramble" in a chain of associations, where self-irony and observation of the world could harmoniously coexist, while never losing sight of his main argument. Rambling is one of the ways of testing the possible.

As I have already said, Hazlitt thought, like Musil, that art's object is the link between the real and the possible. In his essay on *Othello,* he expresses his views about the cognitive value of literature. The critic expands his initial Aristotelian remark about catharsis to include the understanding of literature as an exercise in sympathy and imagination:

> It has been said that tragedy purifies the affections by terror and pity. That is, it substitutes imaginary sympathy for mere selfishness. It gives us a high and permanent interest, beyond ourselves, in humanity as such. It raises the great, the remote, and the possible to an equality with the real, the little and the near. […] It teaches him [the reader, the spectator] that there are and have been others like himself, by showing him as in a glass what they have felt, thought, and done. It opens the chambers of the human heart. It leaves nothing indifferent to us that can affect our common nature. It excites our sensibility by exhibiting the passions wound up to the utmost pitch by the power of imagination or the temptation of circumstances […]. Tragedy creates a balance of the affections. It makes us thoughtful spectators in the lists of life. It is the refiner of the species; a discipline of humanity. ("Othello," Selected Writings vol. 1, p. 112; emphasis added)

Hazlitt's view here of the extended self continues the eighteenth-century debate on sympathy, a debate which is simultaneously ethical and aesthetic. The effect of good fiction, the product of the writer's imagination, is that it sets off the imagination of readers, pushing them to overcome selfishness.

Needless to say, the effect of Shakespeare is the opposite of Rousseau's. The emphasis of the one is quite different from the vividness of the other. As hinted at in Hazlitt's essay, Shakespeare's powerful mind observes the world and the human heart; it brings forth inventive responses from the audience or the readers, and improves their capacity for observation, over a mere projection of the self. In Shakespeare, the passions are "wound up to the utmost pitch," and, as Hazlitt wrote in his essay on *Macbeth:* "All that could actually take place, and all that is only possible

to be conceived, what was said and what was done, the workings of passion, the spells of magic, are brought before us with the same absolute truth and vividness." In Rousseau, passions, on the contrary, are drowned in floods of words:

> His [Rousseau's] fictitious characters are modifications of his own being, reflections and shadows of himself. His speculations are the obvious exaggerations of a mind giving a loose to its habitual impulses…. Hence his enthusiasm and his eloquence, bearing down all opposition. Hence the warmth and the luxuriance, as well as the sameness of his descriptions. Hence the frequent verboseness of his style; for passion lends force and reality to language, and makes words supply the place of imagination. ("On the Character of Rousseau," vol. 2, p. 91)

Like some of the contemporary philosophers mentioned above, Hazlitt suggests here that successful literature is "vivid," offering the possibility of refining the imagination, and that the good writer should provide experiments for the mind and not speculations coming from his "habitual impulses." Even if Hazlitt does not use the term "thought experiment," he is aware of what can challenge engrained habits of thinking. He almost prefigures what Carroll suggests in the aforementioned 2002 article: Thought experiments "help to contemplate, possibly clarify, and even reconfigure our concepts, thereby rendering them newly meaningful." They are ways of proposing distinctions.

According to Hazlitt, the sparkle of experimental thinking is comparison. As stated by Carroll, it should be understood as a device whose aim is not to reach empirical discoveries, but to foster the understanding of conceptual refinements and essential relations. Talking about paintings in "On Imitation," Hazlitt expands once more upon his Aristotelian starting point. Like Aristotle in his *Poetics,* he considers that we take pleasure in mimesis (representation, or imitation to use the more ancient term), and contemplates the same phenomenon Aristotle talked about: The fact that we take pleasure in the representation of unpleasant or even disgusting objects. But Hazlitt adds a new development in his attempts to account for this puzzling pleasure. He actually comes across the primordial emotion with which Aristotle started his *Metaphysics:* that of wonder and astonishment, the cognitive emotion that causes the desire to investigate and to know:

> Objects in themselves disagreeable or indifferent, often please in the imitation. A brick-floor, a pewter-plate, an ugly cur barking […] have been made very interesting as pictures by the fidelity, skill, and spirit, with which they have been copied. One source of the pleasure thus received is undoubtedly *the surprise or feeling of admiration*, occasioned by the unexpected coincidence between the imitation and the object […]. One chief reason, it should seem then, why imitation pleases, is, because, by exciting curiosity, and inviting a comparison between the object and the representation, it opens a new field of inquiry, and leads the attention to a variety of details and distinctions not perceived before. This latter source of the pleasure derived from imitation has never been properly insisted on. ("On Imitation," Selected Writings vol. 2, p. 75; emphasis added)

This passage about the comparison between the represented object and its artistic rendering is a rich description of the indissoluble bond between art, emotions, and cognition.

8.5 Hypothesis in Stendhal

Stendhal's novels are huge thought experiments. *Le Rouge et le Noir,* for example, investigates what happens to an ambitious and intelligent young man of a low social class when he is put in situations of social advancement; *La Chartreuse de Parme* deals with the way in which bright-minded characters operate under the tyranny of a cunning Prince. In both cases, love operates as the discovery of true feelings after a great deal of self-deception. Stendhal's novels meditate on the "just possible," and within them, several narratives develop with descriptions, dialogues, actions, beliefs, desires, and analysis of what has happened in the past and of what might happen in the future. Therefore, several thought experiments occur within the story-telling, which are often marked by the use of hypothetical sentences and the method of nesting imaginings.

These imaginings develop the analysis of a single character. This is most strikingly the case in the analysis of the protagonist of *Le Rouge,* Julien. He observes, through the prism of his own prejudices and his often-wounded social sensibility, the world of the rich, and tries to understand who they are, how they think, and how he can prevail on them. He often reflects in his numerous soliloquies about the way in which he should face that world. Julien is a wonderful specimen of the combination of epistemic virtue and ressentiment, a complex value nuance to which Nietzsche, Steendhal's admirer will return. I will examine one of many examples of this portrait in *Le Rouge et Le Noir.* The day after conquering Mme de Rênal, the wife of the Mayor of the small town of Verrière in whose house he is the tutor of the children, he mistakenly feels offended by an expression in the Lady's face. He believes that she is acting coldly toward him, wrongly attributing what is actually her embarrassment to her social class, thinking that she has the "deliberate intention to put him in his place." He is then upset with himself for his reaction and speculates:

> "Only a fool," he told himself, "loses his temper with other people: a stone falls because it is heavy. *Am I always to remain a boy? When am I going to form* the good habit of giving these people [the rich] their exact money's worth and no more of my heart and soul? *If* I wish to be esteemed by them and by myself, I must show them that it is my poverty that deals with their wealth, but that my heart is a thousand leagues away from their insolence, and is placed in too exalted a sphere to be reached by their petty marks of contempt or favour. (emphasis added)

The narrator specifies in the next paragraph that these musings give to his features "an expression of injured pride and ferocity." But what matters here is the use of the conjunction *if.* The whole sentence is in the conditional mode. The initial questions are, in fact, just the interrogative formulations of Julien's hypothesis: "if I were not a fool, I would not lose my temper," and "if I do not want to remain a boy, I should take the good habit…."

Further, in this same chapter ("A Journey," Book 1, Chap. 12) the misunderstanding between the two lovers grows stronger. Julien takes a journey in the mountains and stops in front of the imposing landscape, an ideal place to forget his agitations in Verrière, and to experience the sublime within nature. He is truly moved by the spectacle of the sublime immensity of the night, and, very quickly, his imagination jumps to his future life:

> In the midst of that vast darkness, his soul wandered in contemplation of what he imagined that he would one day find in Paris. This was first and foremost a woman far more beautiful and of a far higher intelligence than any it had been his lot to see in the country. He loved with passion, he was loved in return. If he tore himself from her for a few moments, it was to cover himself with glory and earn the right to be loved more warmly still.

Conjectures pile up, while indirect speech ("This was first and foremost a woman…") and the hypothetical sentence ("If he tore himself…") portray the triumph of the possible existence Julien imagines leading in Paris.

The dream is interrupted by another series of hypotheses which fill the whole gamut of real, possible, and impossible conditions, as well as conflating the conjectures of several minds:

> Even *if* we allow him Julien's imagination, a young man brought up among the melancholy truths of Paris *would have been aroused* at this stage in his romance by the cold touch of irony; the mighty deeds *would have vanished* with the hope of performing them. (Stendhal 2005, *Le Rouge et le Noir*, p. 412 and 414. From now on, RN. Emphasis added)

The passage sketches a dazzling trip toward the real and the possible. Julien is absorbed in his reverie; the narrator presents his character as differing from his nineteenth-century contemporaries. The novel subtitle is, in fact, *Chronique du XIXe siècle*. The narrator is aware of conjecturing a character, as also of his way of blending features of the historical novel and of the romance. Then there are the young Parisians, cynically aware of the hard realities of existence. Finally, the ambiguous "we" at the beginning of the passage leaves the reader uncertain about the degree of the fiction's verisimilitude (as we will see in the next passage I will quote).

Stendhal here offers one of his quick and dense explorations into the process of imagination. This includes the imagination of people in their everyday conjectures about the future; as mentioned above, de Sousa points out that a whole class of emotions could not exist without the possible, such as fear, hope, doubt, and trust. But Stendhal's narrator also imagines as does Stendhal, in order to stimulate the minds of narrators, readers, and characters.

Stendhal puts forward what is perhaps the most eloquent exploration of the process of imagination in the chapter in which Mathilde, the bright young Parisian aristocrat constantly sensitive to and seeking to realize noble values and virtues, admits to herself that she is in love with Julien. Stendhal interrupts the flow of narration with one of his authorial interventions, introducing reflections about the real, the possible, and the impossible. Appearing in a long paragraph in brackets, this passage shows a complex outburst of various types of irony: playful self-irony; tender irony towards his beloved invention, the character of Mathilde, who is "a sublime soul"; and sharp irony against his contemporaries, whose values are morally low and whose sole passion is self-interest. Tender and satirical moods blend together:

> This page *will* damage the unfortunate author in more ways than one. The frigid hearts *will* accuse it of indecency. It does not offer the insult to the young persons who shine in the drawing-rooms of Paris, of supposing that a single one of their number is susceptible to the mad impulses which degrade the character of Mathilde. This character is wholly *imaginary*, and *is indeed imagined quite apart from the social customs* which among all the ages *will* assure so distinguished a place to the civilisation of the nineteenth century. (RN, 670; emphasis added)

Through the use of the future tense which appears several times in these lines, the narrator imagines what his nineteenth-century reader might imagine: That, for example, *Le Rouge et le Noir* is indecent because it portrays a character who does not correspond to the real people of the century. Of course, the ironic sentence mentioning "the mad impulse which degrades the character of Mathilde" (her falling in love and aspiring to virtue) spells out the opposite of what the writer thinks, as he believes that it is worthwhile to imagine possible characters, since literature, to use the words of Hazlitt in "Othello," refines the species and disciplines humanity.

If the last sentence of the aforementioned lines seems to imply that Stendhal believes, like writers of romances, that fiction is fanciful imagination detached from the real, the following paragraph steers in the reverse direction, showing a realist novel whose ambition it is to mirror reality. This paragraph starts with an imagined answer to a reader who would have protested against the relentless way in which the realist writer reveals people's vices:

> Ah, Sir, *a novel is a mirror carried along a high road*. At one moment it reflects to your vision the azure skies, at another the mire of the puddles at your feet. And the man who carries this *mirror* in his pack *will* be accused by you of being immoral! His *mirror* shows the mire, and you blame the *mirror*! Rather blame that high road upon which the puddle lies, still more the inspector of roads who allows the water to gather and the puddle to form. (emphasis added)

Another final leap seems to confirm the nature of romance but the narrator is actually ironically claiming his right to a constructive utopia:

> Now that it is quite understood that the character of Mathilde is impossible in our age, no less prudent than virtuous, I am less afraid of causing annoyance by continuing the account of the follies of this charming girl. (RN, 671)

Divine irony meanders from mind to mind, from provocation to provocation, and rambles in directions that seem dispersed only to those who do not like to exercise their minds, but actually never loses track of essential truths! The point is that the narrator is contesting the wrong belief that the real and the imaginary are opposites, while proposing that literature should be both: observation of reality and imagination, the real and the possible. Is this not an essential truth about fiction? And about life!

The novel has knowledge value precisely because it investigates the possible as a region of the real; as the contemporary critic D. A. Nuttall says about Shakespeare, the successful writer can "join verisimilitude and wonder" (Nuttall 2007, p. 236). These two aspects recall those Hazlitt alluded to when, in his "On Imitation," he suggested the importance of surprise in the pleasure we take from mimesis. Both Stendhal's hypotheses and Hazlitt's comparisons hint at the possible in art and in life.

Once more, how close they are to Musil! They show that literature can weave together the real, the possible, and the impossible in order to "treat" reality "as a mission and an invention" (Musil 1979, p. 11). How close they are to the speculation of Ulrich in that early chapter of *The Man without Qualities,* where the question of the possible is debated: "Then comes a man who first gives the new possibilities

their meaning and their destiny; he awakens them. […] Since his ideas, insofar as they are not mere idle phantasmagoria, are nothing else than as yet unborn realities, he too of course has a sense of reality."

References

Carroll N (2002) The wheel of virtue: art, literature, and moral knowledge. J Aesthet Art Crit 60/1. http://www.filmschool.lodz.pl/files/show/5-398/CarrolThe+Wheel+of+Virtue.pdf. Accessed 10 Oct 2011

Currie G (2010) Narratives and narrators. Oxford University Press, Oxford

De Sousa R (2005) The art of the possible in life and literature. http://homes.chass.utoronto.ca/~sousa/artpossible.pdf. Accessed 8 Apr 2014

Engel P (2011) Philosophical thought experiments: in or out the armchair? http://www.unige.ch/lettres/philo/enseignants/pe/Engel%202009%20Philosophical%20thought%20experiments.pdf. Accessed 9 Apr 2014

Goldie P (2000) The emotions: a philosophical exploration. Oxford University Press, Oxford

Hartmann N (1953) Aesthetik de Gruyter, Berlin

Hazlitt W (1998) An essay on the principles of human action: being an argument in favour of the natural disinterestedness of the human mind. In: Wu D (ed) The selected writings of William Hazlitt, vol I. Pickering and Chatto, London

Lamarque P (2009) The philosophy of literature. Blackwell, Oxford

Mulligan K (1998) From appropriate emotions to values. The Monist 81:161–188

Musil R (1979) The man without qualities (trans: Kaiser E, Wilkins E). Picador, London

Musil R (1990) The obscene and pathological in art. In: Pike B, Luft DS (eds) Precision and the soul. The University of Chicago Press, Chicago

Nussbaum M (1999) Augustine and Dante on the ascent of love. In: Matthews GB (eds) The Augustinian tradition. California University Press, Berkeley

Nussbaum M (2001) The fragility of goodness: luck and ethics in Greek tragedy. Cambridge University Press, Cambridge

Nuttall AD (2007) Shakespeare the thinker. Yale University Press, New Haven

Reboul A (2001) 'La raison est l'esclave des passions' disait Hume. Actes du Colloque GRAME "L'art, la pensée et les émotions", Lyon. http://hal.archives-ouvertes.fr/docs/00/02/90/59/PDF/Raison.pdf. Accessed 16 Sept 2011

Rescher N (1991) Thought experiments in presocratic philosophy. In: Horowitz T, Massey G (eds) Thought experiments in science and philosophy. Rowman & Littlefield, Lanham

Stendhal (1929) Histoire de la Peinture en Italie I. Le Divan, Paris

Stendhal (1967) Correspondance II. Gallimard, Pléiade, Paris

Stendhal (1968) Correspondance I. Gallimard, Bibliothèque de la Pléiade, Paris

Stendhal (1981) Journal, Œuvres Intimes I. Gallimard, Bibliothèque de la Pléiade, Paris

Stendhal (1982) Souvenirs d'égotisme, Œuvres Intimes II. Gallimard, Bibliothèque de la Pléiade, Paris

Stendhal (1982) Vie de Henry Brulard, Œuvres Intimes. Gallimard, Bibliothèque de la Pléiade, Paris Paris

Stendhal (2005) Le Rouge et le Noir, Œuvres romanesques complètes II. Gallimard, Bibliothèque de la Pléiade, Paris. English version: http://gutenberg.net.au/ebooks03/0300261.txt. Accessed 20 Oct 2011

Chapter 9
L'avenir du Crétinisme

Pascal Engel

«Il nous arrive parfois de penser que le contraire de «crétin» ne serait pas «intelligent», mais «sobre».

Fruttero et Lucentini, Il ritorno del cretino, Mondadori, 1992, tr. fr. Arléa 1993

Devenue Force industrielle, l'Intelligence a été mise en contact et en concurrence avec les Forces du même ordre mais qui la passent de beaucoup comme force et comme industrie.

Charles Maurras, L'avenir de l'intelligence, 1903

Abstract I take up here Kevin Mulligan's idea that stupidity—actually stultitia, foolishness—is insensitivity to the values of knowledge, and argue that this marks the difference between the classical conception and the romantic one. I conclude with some loose considerations about stupidity in contemporary philosophical productions.

Keywords Bêtise · Jugement · Rationalité · Bullshit · Valeurs cognitives

9.1 Le dimanche de l'esprit

On a coutume de dire que l'avenir de la science coïncide avec celui de la bêtise. Au moment même où Renan écrivait sa *Prière sur l' Acropole* et sa célébration de la raison, Flaubert aurait vu que l'envers du décor était la montée du crétinisme. Il est toujours délicat de se livrer à l'histoire rétrospective de la raison et de la déraison, mais il me semble que ce thème – la raison et la science renforcent plus la bêtise que la déraison et l'absence de raison – datent du second romantisme, et appartiennent bien plus à Baudelaire, à Mallarmé, à Valéry et aux auteurs contemporains qu'aux premiers romantiques. Jean-Paul écrivit un *Eloge de la bêtise* en 1781 (Jean-Paul 1993) mais son inspiration est bien plus proche de celle d'Erasme et de Swift que de celle de Flaubert ou Villiers près d'un siècle plus tard. Kevin Mulligan, en musilien,

P. Engel (✉)
Université de Genève, Genève, Suisse
e-mail: pascal.engel@unige.ch

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_9,

a écrit depuis longtemps sur la bêtise, et je n'entends pas ici faire autre chose que d'ajouter une note à ses travaux pionniers (notamment, Mulligan 1998; Mulligan and Engel 2003; Mulligan 2008, 2009).[1]

9.2 Bêtise, intelligence et rationalité

La conception la plus répandue de la bêtise est qu'elle est un défaut intellectuel et cognitif, un manque d'intelligence. Le langage ordinaire désigne celui qui est bête comme un idiot, un crétin ou un imbécile, et de tels termes caractérisent habituellement un défaut de l'entendement ou du jugement. Encore faut-il distinguer le défaut dans la compétence (propre à ceux à qui, comme on dit, il manque une case) et le défaut dans la performance. Une chose est d'être un crétin ou une buse parce que certaines facultés intellectuelles nécessaires à l'intelligence vous font défaut. Autre chose est d'être un idiot parce qu'on manque du jugement comme capacité d'appliquer les règles de l'entendement. On peut avoir les facultés sans avoir le jugement. Kant sanctionne cet usage quand il écrit dans un passage fameux:

> «Le manque de jugement (*Mangel an Urteilskraft*) est proprement ce que l'on appelle stupidité (*Dummheit)*, et à ce vice il n'y a pas de remède. Une tête obtuse ou bornée en laquelle il ne manque que le degré d'entendement convenable et de concepts qui lui sont propres, peut fort bien arriver par l'instruction jusqu'à l'érudition. Mais comme alors, le plus souvent, ce défaut accompagne aussi l'autre, il n'est pas rare de trouver des hommes très instruits qui laissent incessamment apercevoir dans l'usage quils font de leur science ce vice irrémédiable.»[2]

[1] La philosophie de la bêtise – ou plutôt l'essayisme sur la bêtise – est un genre très pratiqué aujourd'hui. Elle donne lieu à des sottisiers, des encyclopédies et à nombre d'essais littéraires ou plus ou moins philosophiques. La plupart sont amusant mais superficiels comme ceux de Canone (2007), ou indigents comme celui de Jerphagnon (2010), qui n'est qu'une rhapsodie de citations qui ne tente même pas de classer les formes de bêtise, ou encore affligeants comme celui de Ronell (2006). Les seuls ouvrages qui ont un intérêt théorique sont: Deleuze (1968); Adam (1975); Rosset (1977); et Roger (2008). Sartre, dans *L'idiot de la famille*, a quelque belles pages sur la bêtise chez Faubert, mais s'englue dans ses catégories dialectico-freudiennes. Sur la bêtise littéraire voir notamment Deshoulières (2005) et Herschberg-Pierrot (2012). Livres et les articles sur la bêtise abondent, sans doute parce que la peur d'être dupe de la bêtise monte autant que le niveau de la bêtise elle-même, que certains même mesurent. De ce point de vue, le livre de Cippola (1988/2012) est un classique de l'approche quantitative et statistique, qui ne tente pas plus de définir la bêtise, mais énonce deux lois fondamentales («Chacun sous-estime toujours inévitablement le nombre d'individus stupides existant dans le monde» et «la probabilité qu'un individu soit stupide est indépendante de toutes les autres caractéristiques de cet individu») Je partage avec Mulligan la conviction qu'on peut *définir* la bêtise et ses espèces, et qu'elle n'est pas aussi insaisissable que les romantiques et le post-modernes le soutiennent. Le vocabulaire de la bêtise est nécessairement «épais» au sens de Bernard Williams, mais cela ne veut pas dire qu'on ne puisse pas distinguer les différences spécifiques au sein de l'espèce: idiot, sot, bête, imbécile, crétin, con, etc. ne désignent pas les mêmes caractéristiques, et celles-ci diffèrent d'une langue à l'autre (*dum*, *moron, nut, tor, sciocco, tonto* ne sonnent pas, d'une langue à l'autre, de la même manière).

[2] KrV tr. Anal. 2. B. Einl. Anm. (I 179—Rc 234) cf K. Eisler, *Kant Lexicon* "Mangel an Urteilskraft". Einem solchen Gebrechen ist nicht abzuhelfen, KrV tr. Anal. 2. B. Einl. Anm. (I 179—Rc 234). Dummheit ist "Mangel an Urteilskraft ohne Witz", Anthr. § 46 (IV 117). Vgl. N 506—523.

Mais que la bêtise tienne au fond (l'entendement et ses concepts) ou à la forme (la capacité à en appliquer les catégories à l'expérience, à bien faire tomber les intuitions sous les concepts), elle demeure, selon cette conception, un déficit intellectuel. *Bouvard et Pécuchet* l'illustre parfaitement. Ils sont deux imbéciles non pas parce qu'ils ne savent rien — au contraire la quantité de savoir qu'ils sont capables d'assimiler est prodigieuse et encyclopédique — mais parce qu'ils ne sont pas capables de l'appliquer. Ils ne savent pas mettre leurs intuitions sous leurs concepts ni adapter leurs concepts à leurs intuitions: en termes kantiens leurs concepts sont vides et leurs intuitions sont aveugles[3].

Cette conception intellectualiste de la bêtise est particulièrement bien représentée au sein de la psychologie contemporaine, qui tend à assimiler la bêtise à un défaut de rationalité manifesté dans l'exercice du jugement. Un grand nombre de travaux de psychologie cognitive et sociale depuis plus d'un demi- siècle se sont employés à montrer que les humains commettent des erreurs systématiques de raisonnement en ne parvenant pas à appliquer des schèmes d'inférence logiques élémentaires. On montre par exemple que les gens échouent à faire des inférences déductives en *modus tollens* (tâche de Wason), qu'ils font des erreurs élémentaires avec le maniement des probabilités, en traitant par exemple la conjonction de la probabilité de deux événements comme plus élevée que celle d'un des conjoints (paralogisme de la conjonction) ou en ignorant systématiquement le taux de base dans des inférences statistiques élémentaires (Kahneman et al. 1982; Wason 1966, et l'immense littérature à laquelle cet essai a donné lieu). Ces erreurs de raisonnement ne sont pas simplement circonstancielles; elles sont profondes et constantes et elles persistent même chez les sujets qui ont reçu une instruction en logique et en statistiques. Le fait que les sujets s'écartent systématiquement des canons de la rationalité déductive (la logique élémentaire) et de ceux de la rationalité inductive et des probabilités (raisonnement bayésien) peut montrer, alternativement, trois choses:

a. les gens sont idiots. Ils sont tout simplement incapables de suivre les normes logiques appropriées;
b. les psychologues sont des idiots. Ils n'ont pas su prendre en compte toutes les variables qui affectent les inférences humaines, et qui, si elles étaient prises en compte, permettraient de montrer que les gens suivent en fait les règles appropriées;
c. Les logiciens sont des idiots. Ils évaluent le comportement logique par rapport à des critères normatifs inappropriés (cf. Thagard 1988, et mon étude de ces options in Engel 1993).

La psychologie contemporaine du raisonnement ne valide évidemment pas la thèse la plus pessimiste (a) qui voudrait que les humains soient *massivement* irrationnels et en ce sens stupides. Mais elle montre au moins, pour reprendre le titre du livre de Paolo Legrenzi, qu'il n'est pas nécessaire d'être stupide pour faire des bêtises, c'est-à-dire des bourdes et des erreurs de raisonnement élémentaires (Legrenzi 2009).

[3] Ceci est très bien vu par Deleuze (1968). Pour prendre la mesure du savoir accumulé par les deux bonhommes, voir notamment Gayon (1998).

Mais ces bourdes ou ces erreurs de raisonnement ne sont pas simplement occasionnelles ou circonstancielles, selon la vision sous-jacente à ces programmes de recherche: elles sont constitutives de l'esprit humain. C'est pourquoi il est un peu naïf de prendre la posture de la raison ou des lumières et de proposer de «réformer le jugement» en corrigeant les erreurs de raisonnement par un meilleur entraînement[4]. Legrenzi donne l'exemple de Bill Clinton dans l'affaire Lewinski. Bien que le président Clinton fût fort intelligent (il ne pouvait prêter au type de soupçons qui atteignaient George W. Bush), il a commis, dans l'affaire Lewinski, plusieurs bourdes notoires: manque d'anticipation, incapacité à évaluer les changements dans la situation (notamment dans les medias et l'usage d'internet), sous -évaluation du risque, trop grande confiance en soi, tendance à prendre ses désirs pour des réalités, et surtout conflit entre la tentation immédiate et les intérêts à long terme. Clinton n'est pas exceptionnel: nous avons tous les mêmes «tunnels mentaux» dans des circonstances différentes qui nous font mal gérer nos choix dans le temps (voir en particulier les travaux de Ainslie 1992, commentés par Elster 1999, et Reach 2005). Selon la conception de la rationalité qui ressort des travaux de la psychologie du raisonnement, les humains subissent des biais systématiques et inévitables aussi bien dans leurs jugements que dans leurs prises de décisions (Cf. Nisbett and Ross 1980; Levit and Lubner 2010; Morel 2002). Selon cette conception les humains sont bêtes parce que leur intelligence subit des limitations nécessaires, qui sont des illusions cognitives, au même titre que sont inévitables les illusions visuelles. Du même coup, si nous sommes stupides, souvent ou quelquefois bien que pas toujours, nous n'y sommes pour rien.

Mais la lecture pessimiste (a) a été elle-même fortement critiquée, au moins sur trois plans. On objecte d'abord que les expériences supposées montrer que nous sommes souvent bêtes dans nos jugements et nos décisions testent en fait les performances des agents, qui sont souvent liées aux circonstances, mais pas leurs compétences, et elles n'invalident pas l'idée selon laquelle les humains ont une compétence rationnelle générale: la rationalité des agents est un trait *a priori* constitutif présupposé par tout test d'intelligence, et non pas une propriété qu'on pourrait infirmer ou confirmer expérimentalement (Cohen 1981; Davidson 1995). On objecte ensuite que tester la rationalité, et en ce sens l'intelligence, est largement dépendant des situations et des circonstances. Une anecdote rapportée par Daniel Dennett l'illustre. L'idiot du village est la risée de la population parce que chaque fois qu'on lui propose le choix entre un *nickel* (cinq cents) et une *dime* (dix cents) il choisit le *nickel*. On lui demande les raisons de son choix et il répond: «Croyez-vous que l'on me re-proposerait ce choix si je choisissais à tous les coups une *dime*?» (Dennett 1991). Enfin, on objecte que les modèles usuels de rationalité, que ce soit en économie en psychologie ou en sociologie, sont inadéquats à représenter la rationalité humaine. Ils sont abstraits, décontextualisés et prêtent aux agents une rationalité maximale, alors que la rationalité humaine est toujours située et contextuelle, limitée et suboptimale. Selon la conception de la rationalité «écologique

[4] C'est la perspective, plutôt *Aufklärer*, de Massimo Piatelli-Palmarini (1998), dans son introduction à cette littérature.

défendue par Gerd Gigerenzer notamment (2009), nombre d'erreurs attribuées par les psychologues de la tradition des «biais et des heuristiques» sont dues à des effets de «cadre» (*framing)* et de contenu auxquels les modèles formels ne sont pas sensibles. Quand on utilise pour les mêmes tâches (par exemple celle qui donne lieu au paralogisme de la conjonction) des modèles différents, comme les modèles fréquentistes pour les probabilités (plutôt que les modèles bayésiens), les soi-disant erreurs diminuent drastiquement. L'intelligence humaine ne procède pas selon des principes généraux tels que les lois logiques, mais selon des «heuristiques rapides et frugales» qui permettent d'économiser des ressources finies et situées. Si on adapte ainsi nos modèles de rationalité, les gens apparaissent bien moins idiots qu'ils n'en ont l'air. Le paradoxe de la conception écologique de la rationalité est qu'elle est tout autant une conception de l'intelligence que de la bêtise: nos bêtises sont adaptatives et si nous faisons des erreurs c'est parce qu'il nous faut optimiser notre comportement à notre environnement. Nous sommes donc nécessairement bêtes, et c'est, dans l'ensemble, une bonne chose. Ce serait le contraire qui serait inquiétant.

Mais que l'on adopte la conception normative ou la conception écologique de la rationalité, l'intelligence ou la bêtise demeurent conçues comme des compétences ou absences de compétences *intellectuelles* et *générales* s'exerçant dans des circonstances particulières de jugement et de décisions. Ainsi tout le monde les a, ou manque à les avoir, en partage, et chacun a sa part d'intelligence ou de bêtise, selon les circonstances. C'est une conception très démocratique de la bêtise, illustrée parfaitement par la devise de cet autre parangon de démocratie intellectuelle, Forrest Gump: «*Stupid is as stupid as stupid does*» – tout un chacun a sa part de stupidité. Il n'y a pas besoin d'être stupide pour faire des bêtises, et il n'y a pas besoin d'être intelligent pour se comporter intelligemment. Selon cette conception démocratique, la bêtise n'est pas une propriété de certains individus et de certains caractères, c'est une propriété de certains actes et de certains jugements, et il n'y a pas de *personnes* stupides qui auraient la bêtise en partage: elle est parfaitement partagée, comme le bon sens selon Descartes. Ce n'était ni la conception classique, ni la conception romantique de la bêtise.

9.2.1 Bêtise classique et bêtise romantique

Ce que j'appelle «bêtise classique» et «bêtise romantique» sont plus des idéaux-types que des catégories historiques bien définies. Il ne s'agit pas tant de se réclamer d'une histoire que d'une manière de comprendre ces notions. Les classiques et les romantiques considèrent la bêtise comme un défaut non pas de rationalité ou de jugement comme les contemporains, mais comme un défaut de *raison*, la raison n'étant pas seulement la faculté de raisonner, mais celle qui obéit à des principes et surtout à des valeurs, qui sont non seulement des valeurs morales (la justice, le bien), mais aussi des valeurs esthétiques (le beau) et surtout des valeurs cognitives (le vrai). Mais là où le classique voit dans la bêtise – ou plus exactement dans la

sottise – un envers calamiteux de la raison, le romantique se laisse fasciner par cet envers.

La bêtise selon les classiques n'est pas d'abord un trait de certaines actions ou de certains jugements, mais de certains individus et de certains caractères. Elle est une certaine sorte d'inhérence, une propriété essentielle de celui qui l'a, et non pas un accident. Comme le dit la Bruyère (1951, p. 359), «La sottise est dans le sot, la fatuité dans le fat, et l'impertinence dans l'impertinent». Ce qui est bête c'est la personne tout entière. Les classiques, en second lieu, ne caractérisent pas la bêtise comme étant seulement un défaut intellectuel ou un manque de jugement, mais avant tout comme un défaut de la *sensibilité*. C'est notamment pourquoi ils emploient très rarement le terme de *bêtise*, mais plutôt celui de *sottise*. Le sot n'est pas bête seulement au sens où il manquerait de jugement, de logique, ou d'intelligence, mais parce qu'il a un défaut de la sensibilité. De la sensibilité aux autres d'abord: le sot est un vaniteux, celui qui est plein de son moi. De la sottise à la fatuité il n'y a qu'un pas. A nouveau La Bruyère: «Un fat est celui que les sots croient homme de mérite» (1951, p. 358). Ou encore chez La Fontaine ce mulet vaniteux qui se vantait de sa généalogie, et à qui il arrive que «étant devenu vieux on le mit au moulin».

> Quand le malheur ne serait bon
> Qu'à mettre un sot à la raison,
> Toujours serait-ce à juste cause
> Qu'on le dit bon à quelque chose.»

Un sot mis à la raison n'est pas quelqu'un qu'on rend plus intelligent ou auquel on donne de l'esprit. Car il ne suffit pas d'avoir de l'esprit pour échapper à la sottise. Il faut aussi avoir un certain *respect* pour les *valeurs* de l'esprit, et au premier chef pour la vérité (Sur ce point voir le bel essai de Johansson 2006). Ou plus exactement, avoir *réellement* de l'esprit c'est aussi avoir un sens des *valeurs* de l'esprit. C'est pourquoi les classiques distinguent quelqu'un qui a de l'esprit – donc qui n'est pas un sot – de celui qui est un *bel esprit*. Le bel esprit est celui qui prétend avoir de l'esprit, et en a sans doute une part, mais qui ne respecte pas la vérité parce qu'il ne songe qu'à briller. Rien n'est plus parlant à cet égard que le mot de Malebranche qu'il faut mettre au fronton de toute théorie de la bêtise:

> «Le stupide et le bel esprit sont également fermés à la vérité; il y a toutefois cette différence que le stupide esprit la respecte tandis que le bel esprit la méprise» (Malebranche 1992b, p. 671).

Le stupide manque de jugement et ne peut accéder au vrai. Mais le bel esprit *méprise* la vérité comme valeur et n'en a cure. Par là il rejoint celui qu'Erasme appelait le *fol*, et les latins le *stultus*: *Stultitia* et *moria* sont les noms latins de la sottise, qui n'est pas du tout l'absence d'intelligence, mais l'absence de sagesse, propriété morale et non pas intellectuelle (Chamfort: «La plupart des folies ne viennent que de sottises»). Chez Sénèque le *stultus*, le sot, est celui qui s'agite sans raison et s'occupe sans cesse de choses vaines et sans intérêt, qui n'est pas maître de son

temps[5]. Chez les grands moralistes, le sot est celui qui manque de valeurs morales (La Rochefoucauld: «Un sot n'a pas assez d'étoffe pour être bon», *Maximes* 387, 1964, p. 454), mais surtout de valeurs intellectuelles. Il est simplement aveugle à celles-ci. Pour les penseurs classiques, la raison et la vérité sont des propriétés objectives, et les valeurs correspondantes sont tout aussi objectives. Celui qui n'est pas capable d'avoir les attitudes appropriées, l'esprit et la véracité, qui sont des vertus cognitives ou intellectuelles[6], est quelqu'un à qui il manque non seulement une certaine compétence cognitive, mais aussi une certaine compétence affective et par là même éthique.

C'est une tout autre perspective que vont adopter les romantiques. C'est avec le romantisme que la bêtise devient un thème littéraire et philosophique de plein droit. Jean Paul écrit un *Eloge de la bêtise* qui parodie Erasme, mais porte sur tout autre chose que la *stultitia* et la sottise des classiques. Il y a deux traits principaux de la bêtise romantique (Voir Roger 2008, Chap. II). Le premier est que la bêtise est l'envers de la raison classique, donc des valeurs du vrai, mais à la différence du classique, qui la méprise, le romantique est fasciné par elle, par le type de fascination qui s'allie au mépris. C'est l'envers du bien juger kantien: alors que bien juger, donc avoir de l'esprit, selon Kant, c'est bien appliquer les catégories au divers de l'intuition sensible, celui qui est bête et juge mal, selon le romantique, est celui qui produit des jugements au-delà des catégories. Cela correspond en fait à la définition kantienne du sublime. Il y a une sourde affinité entre le sublime comme catégorie esthétique et la bêtise, que Schopenhauer a très bien vue, ce qui fait de lui le grand penseur de la bêtise du XIXème siècle romantique. Pour Schopenhauer, la bêtise est une inaptitude à faire usage du principe de raison suffisante, qui nous fait sortir de la représentation et nous livre à la volonté qui égalise les plus grands génies et les met sur le même plan que les animaux (brutes) et les plantes (légumes): elle met tout le monde, comme le dit Deleuze qui reprend la substance des vues de Schopenhauer, sur le même fond «digestif et légumineux» (Deleuze 1968, cité par Roger 2008, p. 28). Deleuze assimile la bêtise à ce fond sans fond abysssal dont il voit chez Schelling l'incarnation. Peut-être était quelque chose de ce genre que voulait dire Renan quand il disait qu'il n'y a qu'une seule chose qui donne l'idée de l'infini, à savoir la bêtise humaine (Renan 1876). Le second grand trait de la bêtise romantique est qu'elle est le propre des foules et des masses. Pour le romantique la bêtise n'est pas l'attribut de quelques-uns, mais de tout le monde. C'est la foule qui est bête, qui égalise tout et digère tout. Cette thématique court de Poe et Baudelaire à Flaubert, et de Nietzsche à Musil et Valéry. La foule bête s'oppose au génie et au poète mais elle le fascine aussi: le poète et l'idiot ne contemplent-ils pas tous deux ce puits sans fond qu'est la stupidité humaine? Les deux thèmes, celui de la bêtise

[5] Humilis res est stultitia, abiecta, sordida, seruilis, multis affectibus et sacrissimis subiecta. Hos tam graues dominos, interdum alternis imperantes, interdum pariter, dimittit a te sapientia, quae sola libertas est. Una ad hanc fert uia, et quidem recta; non aberrabis; uade certo gradu. Si uis omnia tibi subicere, te subice rationi; multos reges, si ratio te rexerit. Ab illa disces quid et quemadmodum aggredi debeas; non incides rebus (à Lucilius, IV, 37).

[6] Ce que Bernard Williams (2002/2006) appelle les vertus de vérité sont les vertus classiques par excellence, même si comme Williams le montre elles ont une histoire.

comme anti-raison et celui de la bêtise comme propre des foules se conjuguent dans l'idée, elle aussi très romantique, selon laquelle la bêtise est l'incarnation des principes mêmes de la raison et avant tout celui d'identité. L'imbécile, de Prudhomme à Homais, est celui qui énonce des tautologies, des proverbes, des idées reçues et des lieux communs, et qui épuise la raison dans l'usage du principe d'identité: «Les affaires sont les affaires», «tous les Juifs sont des Juifs», «Vive la Pologne, car s'il n'y avait pas de Pologne il n'y aurait pas de polonais!» etc. La sagesse des nations devient la sottise des peuples[7]. Ce que le romantique – ou le post romantique comme Flaubert, hait dans la bêtise est exactement le contraire de ce que le classique lui reprochait: le fait qu'elle rejoigne, par sa célébration du principe d'identité A = A, l'essence de la raison. C'est la raison elle-même qui est «conne» (Picard 1994, p. 54). Comme le dit Deleuze, résumant parfaitement le retournement des Lumières au romantisme et prononçant un credo qui sera aussi celui des post-modernes: «Ce n'est pas le sommeil de la raison qui engendre des monstres, mais la rationalité vigilante et insomniaque» (Deleuze and Guattari 1972, p. 133). D'où un autre thème, que l'on retrouvera jusqu'à Bergson et à l'existentialisme: la bêtise c'est la raison statique et immobile, avec ses principes, ses concepts vides et ses vérités permanentes, alors que ce qui s'oppose à elle c'est le dynamique et le temporel qui sont source d'évolution créatrice et qui ne se saisissent que dans l'intuition[8].

J'ai brossé à grands traits trois conceptions de la bêtise: la conception classique selon laquelle elle est aussi bien un défaut intellectuel qu'un vice moral – une insensibilité aux valeurs cognitives-, la conception romantique selon laquelle c'est l'incarnation de la logique et de la raison identitaire qui écrase les différences et la créativité, et la conception contemporaine démocratique selon laquelle elle est la manifestation d'une irrationalité partagée par l'espèce humaine. Laquelle est correcte? La réponse est difficile, parce que, comme l'on remarqué tous ceux qui se sont attaqués à la bêtise, le phénomène est pluriel et fuyant. Dans la *Dunciade* de Pope, la bêtise (*dullness*) se présente sur son trône comme une reine ou une déesse dotée de tous ses attributs – foulant à ses pieds la science, l'esprit, la logique et la morale – mais la bêtise moderne et contemporaine est multiforme. Selon le credo contemporain, peut-être n'y a-t-il pas une essence de la bêtise, et peut-être est-elle affaire de contextes, d'occasions: tout le monde est un peu bête tout le temps. Peut-être chacune des conceptions que j'ai distinguées offre-t-elle une facette de cette propriété multiforme. La conception romantique est celle qui domine. Pourtant, s'il y a une conception qui a le plus de chances d'être correcte, c'est bien la conception classique. L'erreur commune à la conception romantique et à la conception contemporaine consiste à voir dans la bêtise un défaut principalement intellectuel – manque

[7] Alain Roger, qui défend nettement la conception romantique de la bêtise, fait de la tautologie l'essence de la bêtise, reprenant une idée de Schopenhauer qui sera aussi au centre des essais de Rosset (1977).

[8] Benda, mieux que tout autre, dans sa hargne contre le bergsonisme, vit parfaitement comment ce dernier récupéra les notions de vie, de changement et de dynamisme au bénéfice de l'intelligence et de l'intuition, opposées au statisme du concept et de la raison. Pour Benda, la bêtise, c'est Belphégor, le dieu amorphe. Pour Bergson, la bêtise c'est le contraire de l'intelligence créatrice et dynamique.

de logique ou excès de logique – alors que la bêtise est aussi un défaut de la sensibilité et de l'affect, et plus exactement une incapacité de l'affect de se mettre en harmonie avec les exigences de l'entendement et de la raison. Personne, mis à part les grands moralistes classiques, n'a mieux vu cela que Robert Musil. Dans son célèbre essai *Uber die Dummheit*, il distingue la bêtise honnête, celle du benêt ou du niais, de la bêtise prétentieuse, supérieure, qui est

> «moins un manque d'intelligence qu'une abdication de celle-ci devant les tâches qu'elle prétend accomplir alors qu'elles ne lui conviennent pas; elle peut comporter tous les caractères négatifs d'un entendement faible, mais avec en plus, tous ceux qu'impliquent une affectivité déséquilibrée, contrefaite, irrégulière, en un mot: maladive. Comme il n'a pas d'affectivités «normalisées», cette déviation maladive traduit plus précisément une dysharmonie entre les partis pris du sentiment et un entendement incapable de les modérer. Cette bêtise supérieure est la vraie maladie de la formation…Elle peut affecter la plus haute intellectualité.» (Musil 1984, pp. 314–315).[9]

C'est exactement cette bêtise «intelligente», comme l'appelle Musil, que Malebranche diagnostiquait chez le «bel esprit» qui méprise la vérité. Elle peut, comme le remarque Musil, faire usage de toutes les vérités et s'en emparer, «prendre tous les costumes de la vérité», alors que «la vérité, elle, n'a jamais qu'un seul vêtement, qu'un seul chemin: elle est toujours handicapée.» Que la bêtise puisse prendre les costumes de la vérité, c'est justement là son leurre suprême.

Comme le disait Bacon:

> Lorsque l'erreur porte les livrées de la vérité, elle est souvent plus respectée que la vérité même, et ce faux respect a des suites très dangereuses. *Pessima res errorum apotheosis, et pro pester intellectus habenda est, si vanis accedat venaratio.* (Bacon, *Novum organum* I)

9.3 L'empire de la foutaise

On a affaire, dans nombre de ces circonstances (l'affaire Sokal étant un paradigme) très exactement à ce que Harry Frankfurt (2005) appelle, dans un essai qui a eu un immense succès, *On Bullshit*, à de la *foutaise*[10]. La foutaise, nous explique-t-il est un phénomène extrêmement répandu dans notre culture. Produire, par des articles, des livres, des interviews de journaux, et aujourd'hui encore plus massivement que jamais dans l'histoire de l'humanité, avec internet, sur des blogs, des sites variés, de la foutaise, ce n'est pas mentir, ou déroger au vrai, au sens où l'on ferait des erreurs, des jugements faux ou même où l'on ferait des mensonges. Comme le dit Frankfurt, celui qui dit de la foutaise n'est

> pas en train d'exprimer un énoncé qui serait vrai ou faux, comme un mensonge: L'essence de la foutaise est simplement un manque de connexion avec un souci *(care)* pour la vérité – une indifférence à la question de savoir ce qu'il en est réellement.

[9] Presque tout ce qu'il y à dire ici l'est par Kevin Mulligan, *op cit*.

[10] Le titre de la traduction française est *L'art de dire des conneries*, qui ne me paraît pas rendre l'idée.

Le *bullshiter* est littéralement quelqu'un qui *se fout* de dire quoi que ce soit de vrai ou de faux et *se fout de nous*. Il n'a aucun respect pour la vérité, ni pour les valeurs cognitives. Il se moque de dire des choses vraies, justifiées, confirmées, ou informées. Il se moque du fait que ce qu'il dit de la science, de la philosophie ou des œuvres de l'esprit soit correct ou pas. Ce qui l'intéresse c'est seulement d'en dire quelque chose, et si possible quelque chose qui soit nouveau, intéressant, curieux. Il est, dans un univers auquel les medias ont donné une chambre d'écho inégalée, la réincarnation du bel esprit «fermé à la vérité» dont parlait Malebranche. Ce dernier avait d'ailleurs parfaitement diagnostiqué l'une des passions du bel esprit et du producteur de foutaise: un usage mal placé de la *curiosité*, qui devrait être une vertu intellectuelle, mais qui entre leurs mains devient un vice:

> «Les savants mêmes et ceux qui se piquent d'esprit passent plus de la moitié de leur vie dans des actions purement animales. Ils font de leur tête une espèce de garde meuble, dans laquelle ils entassent sans discernement et sans ordre…Ils se font gloire de ressembler à ces cabinets de curiosités et d'antiques, qui n'ont rien de riche ni de solide, et dont le prix ne dépend que de la fantaisie, de la passion et du hasard, et ils ne travaillent presque jamais à se rendre l'esprit juste» (*Recherche de la vérité,* Préface, in *Œuvres, I*, 1992)

Dans la science, l'auteur post-moderne dont se moque Sokal aime ce qui est nouveau, ce qui surprend. Tous ceux qui lisent les ouvrages et revues de popularisation scientifique le savent: la mécanique quantique, même si on n'y comprend rien, est bien plus intéressante que la mécanique classique (qui réserve pourtant bien des surprises); les structures dissipatives sont bien plus drôles que les structures assimilatives, le théorème de Gödel bien plus intéressant que celui de Löwenheim-Skolem (qui n'est pas si trivial), les machines de Turing bien plus rigolotes, avec leurs rubans, que la plate thèse de Church (qui est plus puissante), la théorie des catastrophes bien plus amusante que la dynamique des fluides (qui est très complexe), la logique dynamique bien plus porteuse que la barbante logique modale (qui peut être tout aussi dynamique), la vie artificielle bien plus sexy que la vie naturelle (alors que la première est au contraire ennuyeuse), etc. Dans la philosophie, l'herméneutique est bien plus excitante que la bête histoire des idées, la lecture de Derrida sur Searle est bien plus drôle que l'inverse, etc. En littérature Christine Angot est plus intéressante que Paul Bourget et Amélie Nothomb que George Sand (alors que je ne vois pas trop la différence). Un Bouvard d'aujourd'hui aurait beaucoup à surfer.

Les auteurs qu'attaquaient Sokal et Bricmont ne sont donc pas tant des idiots qu'ils ne sont des sots, des *fols*, au sens le plus classique du terme. Ils sont remarquablement intelligents, compétents, instruits, et même, à bien des égards, savants. Ce sont tous des universitaires de haut vol, des érudits, des intellectuels remarquablement subtils et sophistiqués, auteurs d'ouvrages complexes et, à bien des égards, remarquablement inventifs. Mais ils sont typiquement des cas de ce que Musil appelle la «bêtise intelligente».

Ce que Sokal voulut renverser, c'est précisément cette *audace*, cet *héroïsme* que l'on célèbre tant dans la pensée contemporaine. Depuis plus d'un siècle, depuis en fait Bergson, on célèbre la nouveauté, l'inventivité de la pensée française contemporaine, son caractère *osé*. On nous a parlé de l'aventure du bergsonisme, de

l'existentialisme, du structuralisme, de la nouvelle philosophie, en décrivant les exploits de ces penseurs comme des «aventures».[11] La pensée de la foutaise est une pensée qui ose, qui a un sacré culot. Mais oser, ce n'est pas simplement proposer, comme on le dit souvent, une avancée, une «percée» dans la pensée. C'est aussi manquer de respect pour des valeurs, des interdits, des normes. Ici celles de l'intellect. Michel Audiard disait que les cons ça ose tout, et que c'est même à cela qu'on les reconnaît. Ce n'est pas complètement vrai. Certains osent seulement dans le domaine de l'intellect.

9.4 La bêtise philosophique

Les philosophes, de Malebranche à Deleuze, adorent dénoncer la bêtise et la diagnostiquer. La nouveauté est qu'ils en sont devenus, en ce début de vingt-et unième siècle, parmi les principaux producteurs. G.A. Cohen a suggéré que si les Français avaient acquis, dans ce domaine une sorte de monopole, c'est parce qu'ils enseignent la philosophie dans l'enseignement secondaire (Cohen 2002). L'argument n'est pas très bon, car les Etats-Unis, pour ne prendre que leur cas, sont parvenus à créer autant de *bullshit* que la plupart des autres pays réunis alors même qu'on n'y enseigne pas la philosophie dans le secondaire et qu'elle est une discipline académique très minoritaire dans la culture des medias, à la différence de ce qui se produit en France, en Italie ou en Espagne. Le même genre de jugements que celui de G. A. Cohen ont été portés par Max Black, Bernard Williams et Galen Strawson au sujet de la production philosophique de langue anglaise, quand ils ont déclaré que quasiment 98 % de la production philosophique anglophone était nulle. Alors qu'est-ce qui fait la spécificité française? La nosologie philosophique doit aussi se compléter de l'étiologie[12].

Dans l'un des premiers essais de nosologie philosophique du vingtième siècle, *Belphégor* (1918, p. 171), Julien Benda donnait comme cause la perte des valeurs intellectuelles, qu'il identifiait au règne du romantisme et de l'esthétisme, le fait qu'elle «soit tout entière faite par les femmes» parce que ce sont elles qui ont des loisirs et le temps pour des activités de luxe, dont la philosophie fait partie. Cette explication n'est pas bien bonne, car les hommes ont abondamment fait la preuve de leur capacité à user de leurs loisirs de la manière que Benda réprouvait, et il faut bien dire que les principaux philosophes capables d'incarner la bêtise philosophique en France sont de sexe masculin, les femmes philosophes devant se contenter, non sans de brillantes exceptions, de la portion congrue. Les sociologues de l'école de Bourdieu se sont aussi attelés à la tâche, et ils ont produit des hypothèses

[11] Alain Badiou (2005) a parlé de «l'aventure de la pensée française contemporaine et un livre porte le nom: «Vincennes: une aventure de la pensée française: Vincennes» (Djian 2009).

[12] Parmi d'excellentes nosologies, il y a les essais de Frederic Nef et de Kevin Mulligan que j'ai jadis publiés dans *Stanford French Review* 17. 1994, mais personne ne semble les avoir lus. L'article de Jon Elster (2011) dresse une bonne nosologie, mais ne donne aucune étiologie.

convaincantes, comme la progressive médiatisation d'une bonne partie du champ intellectuel (Pinto 2007). Les noces de la philosophie et du journalisme ne datent pas, en France, d'hier. Alain, puis Sartre avaient préparé le terrain. Michel Foucault proposa explicitement de définir la philosophie comme une «ontologie du présent». Aujourd'hui la philosophie médiatique est un marché, qui a ses animateurs vedettes, ses émissions phares à la radio et à la télévision, ses magazines, ses *groupies* et ses *geeks*. Le fait qu'elle pèse sur l'éducation, tende à définir les sujets du bac et à imposer ses thématiques – qui relèvent toutes peu ou prou de l'éthique, et qui excluent par principe les parties les plus théoriques de la philosophie («Cela n'intéresse personne») – tend à donner raison à G.A. Cohen. La situation française ressemble certes à celles d'autres pays, et la philosophie médiatique se porte bien partout en Europe, de l'Espagne à la Finlande, de l'Angleterre au Portugal et à l'Italie Mais la massification entraîne-t-elle par elle-même la bêtise? Elle ne l'entraîne que si les conditions et les normes de la pensée changent et deviennent celles du journalisme: n'accepter un argument que s'il produit des effets, une pensée que si elle est frappante, une thèse que si elle est d'actualité, ignorer systématiquement tout ce qui peut coûter un effort intellectuel. Le succès de la philosophie dans les medias et la vie culturelle en général accentue un phénomène qui est l'une des marques de l'absence de respect pour les valeurs intellectuelles: le faux raisonnement ou le pseudo raisonnement (*sham reasoning*).[13] Un autre facteur important est l'absence, en France en particulier, de véritables critiques dans les journaux: la plupart des recensions d'ouvrages sont des panégyriques ou des passages de brosse à reluire, et quand un article est un peu critique ou négatif, on crie à la polémique ou à la malveillance. Il n'y a pratiquement pas en France d'équivalent des rubriques *Letters to the Editor* des journaux anglais ou américains. Tous ces traits de la culture journalistique contribuent à la marche invicible de la sottise. Ce qui produit la sottise c'est la généralisation de ce que Malebranche décrivait comme le comportement des beaux esprits, sous l'espèce de ce qu'on appelait pas encore le snob:

> «Un mauvais mot, un accent de province, une petite grimace les irrite infiniment plus qu'un amas confus de méchantes raisons. Ils ne peuvent reconnaître le défaut d'un raisonnement, mais ils sentent parfaitement bien une fausse mesure et un geste mal réglé. En un mot ils ont une parfaite intelligence des choses sensibles, parce qu'ils ont fait un usage continuel de leurs sens; mais ils n'ont point la véritable intelligence des choses qui dépendent de la raison, parce qu'ils n'ont presque jamais fait usage de la leur.» (*Recherche de la vérité*, II ii, 8, *op cit*)

[13] Cf. Peirce (1931/1947, p. 56) "Lessons from the history of science": The effect of mixing speculative inquiry with questions of conduct results finally in a sort of half make-believe reasoning which deceives itself in regard to its real character.… In short, it is no longer the reasoning which determines what the conclusion shall be, but it is the conclusion which determines what the reasoning shall be. This is sham reasoning. In short, as morality supposes self-control, men learn that they must not surrender themselves unreservedly to any method, without considering to what conclusions it will lead them. But this is utterly contrary to the single-mindedness that is requisite in science.» Peirce avait bien vu le rapport entre le mauvais raisonnement et l'abaissement de la moralité.

Voir aussi Paulhan (1889).

Mais aucune explication n'est plus simple que celle que produisit Benda (1945, p. 149) dans *La France Byzantine*: «le snobisme a pris des proportions qu'on ne lui avait encore jamais vues; il est devenu l'opinion.»

References

Adam M (1975) Essai sur la bêtise. PUF, Paris
Ainslie G (1992) Picoeconomics. Cambridge University Press, Cambridge
Badiou A (2005) The adventure of French philosophy. N Left Rev 35:67–77
Benda J (1918) Belphegor. Emile Paul, Paris
Benda J (1945) La France Byzantine. Gallimard, Paris
Canone B (2007) La bêtise s'améliore. Stock, Paris
Cippola C (1988) The basic laws of stupidity. Il mulino, Bologna (French trans: Paris, PUF 2012)
Cohen LJ (1981) Can human rationality be experimentally demonstrated? Behav Brain Sci 4(3):317–331
Cohen GA (2002) Deeper into bullshit. In: Buss S, Overton L (eds) Contours of agency: essays on themes from Harry Frankfurt. MIT Press, Cambridge, pp 322–339
Davidson D (1995) Can there be a science of rationality? In: Problems of rationality. Oxford University Press, Oxford
Deleuze G (1968) Différence et répétition. PUF, Paris
Deleuze G, Guattari F (1972) L'anti oedipe. Minuit, Paris
Dennett D (1991) La stratégie de l'interprète. Gallimard, Paris
Deshoulières V (2005) Métamorphoses de l'idiot. Klincksieck, Paris
Djian JM (2009) Vincennes: une aventure de la pensée française: Vincennes. Flammarion, Paris
Elster J (1999) Alchemies of the mind. Cambridge University Press, Cambridge
Elster J (2011) Obscurantisme dur et obscurantisme mou dans les sciences sociales. Diogène 229(30):231–247
Engel P (1993) Logique, raisonnement et rationalité. In: Houdé O (ed) Pensée logico mathématique. PUF, Paris
Frankfurt HG (2005) On bullshit. Princeton University Press, Princeton
Gayon J (1998) Agriculture et agronomie dans *Bouvard et Pécuchet* de Gustave Flaubert. Littérature 109:59–73
Gigerenzer G (2009) Penser le risque. Markus Haller, Geneve
Herschberg-Pierrot A (ed) (2012) Flaubert, l'empire de la bêtise. Editions Cécile Default, Nantes
Jean-Paul (1993) Eloge de la bêtise (French trans: Briand N), préface de Hermann Hesse. Corti, Paris
Jerphagnon L (2010) La sottise. Albin Michel, Paris
Johansson I (2006) Respect for logic. University of Göteborg, web series 36
Kahneman D, Slovic P, Tversky A (eds) (1982) Judgment under uncertainty: heuristics and biases. Cambridge University Press, Cambridge
La Bruyère J (1951) Caractères. Gallimard, La Pléiade, Paris
La Rochefoucauld F (1964) Maximes. Galimard, Pléiade, Paris
Legrenzi P (2009) Non occore essere stupidi per fare schiochezze. Il Mulino, Bologna
Levit SJ, Lubner SD (2010) Freakonomics. Gallimard, Paris
Malebranche N (1992a) Recherche de la vérité, Œuvres, I. Gallimard, La Pléiade, Paris
Malebranche N (1992b) Entretiens sur la métaphysique et la religion, Œuvres, II. Gallimard, La Pléiade, Paris
Morel C (2002) Les décisions absurdes. Gallimard, Paris
Mulligan K (1998) Valeurs cognitives. Le Magazine Littéraire 361:78–79
Mulligan K (2008) Ironie, *valeurs cognitives* et bêtise. Philosophiques 35(1):89–107

Mulligan K (2009) Torheit, Vernunftichkeit und der Wert des Wissens. In: Schönrich G (ed) Wissen und Werte. Mentis, Padeborn, pp 27–44
Mulligan K, Engel P (2003) Normes éthiques et normes cognitives. Cités 15(3):171
Musil R (1984) Uber die Dummheit. Vortrag auf Einladung des österreichischen Werkbunds Gehalten in Wien am 11. März und wiederholt am 17. März 1937. In von Frisé A (Hrsg) Gesammelte Werke (Bd IV). Rowohlt, Reinbe (Essays und Reden Tr. fr P, Jacottet In: Essais. Seuil, Paris)
Nisbett RE, Ross L (1980) Human inference: strategies and shortcomings of social judgment. Prentice-Hall, Englewood Cliffs
Paulhan F (1889) Esprits logiques et esprits faux. Alcan, Paris
Peirce CS (1931/1947) Collected papers, vol I. Harvard University Press, Cambridge
Piatelli-Palmarini M (1998) La réforme du jugement ou comment ne plus se tromper. O. Jacob, Paris
Picard G (1994) De la connerie. Corti, Paris
Pinto L (2007) La Vocation et le métier de philosophe. Pour une sociologie de la philosophie dans la France contemporaine. Seuil, Paris
Reach G (2005) Pourquoi se soigne-t-on? Le bord de l'eau, Bordeaux
Renan E (1876) Dialogues et opinion philosophiques. Calman Levy, Paris
Roger A (2008) Brévaire de la bêtise. Gallimard, Paris
Ronell A (2006) Stupidity. Stock, Paris
Rosset C (1977) Le réel, Traité de l'idiotie. Minuit, Paris
Thagard P (1988) Computational philosophy of science. MIT Press, Cambridge
Wason PC (1966) Reasoning. New Horiz Psychol 1:135–151
Williams B (2002) Truth and trutfulness. Princeton University Press, Princeton. French edition: Williams B (2006) Vérité et véracité. Gallimard, Paris

Part II
Epistemology, Perception, and Consciousness

Chapter 10
Three Easy Points on Relative Truth

Diego Marconi

Abstract As a contribution to the debate on the intelligibility of the notion of relative truth**,** I discuss three issues that are of some interest in the way of bush-beating. They are (1) whether relative truth can be explicated as truth in a subjective world, (2) whether alleged relative truth could just be belief (i.e., "*p* is true for *X*" = "*X* believes that *p*"), and, finally (3) whether plain truth could, or should be defined on the basis of relative truth. The first two questions receive a negative answer, while the third is seen to depend on further decisions on the nature of relative truth, though one particular attempt at articulating the relation between plain and relative truth (Kölbel, Relative truth, 2002) is shown to be unconvincing.

Keywords Facts · Relativism · Truth · Faultless disagreement · Judgments of taste

10.1 Preliminaries: Moderate Relativism, Faultless Disagreement, "True for X"

Here, I will use the phrase "moderate relativism" as it is used by Crispin Wright (2008), i.e., to describe the view on which relativistic accounts of the truth conditions of propositions are appropriate for some propositions, though not for every proposition as in radical, or global, or universal relativism (Wright 2008).[1] We say that a semantic account is relativistic (as distinct from a contextualistic account) if the truth value of a contextually complete propositional content is made to depend not just on a possible world, but on an extra parameter as well, the extra parameter being intuitively interpreted as a perspective (Kölbel 2002), or a judge (Lasersohn 2005), or a point of assessment (MacFarlane 2005). As radical relativism is widely believed to be shot through with insurmountable difficulties (e.g., Boghossian 2006), nowadays philosophers interested in relativism incline to the more defensible moderate variety.

[1] "Moderate relativism" is used in a different sense both in Recanati (2007) and in Lopez de Sà (2008).

D. Marconi (✉)
University of Turin, Turin, Italy
e-mail: diego.marconi@unito.it

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_10,

The view that relativistic semantics applies to certain classes of propositions (though not to all) has been supported by the widespread intuition that simple, "monadic" truth (Cappelen and Hawthorne 2009) is not at home in some domains of discourse: For example, judgments of taste cannot plausibly be said to be unqualifiedly right or wrong, and the propositions they express are not plausibly regarded as simply true or simply false, as there appear to be no "facts of the matter" to decide issues of taste. If Ann asserts:

(1) (*Ann*) Korean cabbage is delicious,

while Bill retorts:

(2) (*Bill*) No, it is not: it is disgusting,

their disagreement (for they do disagree, or so the relativistic intuition goes) does not entail that either Ann or Bill must be wrong as either assertion must be false. Their disagreement is *faultless* (Kölbel 2002): they may be both right, in a sense. Relativistic accounts of the truth conditions of the propositions expressed by judgments of taste are meant to capture the "sense" in which they may be both right.

Moderate relativism is committed to the possibility of faultless disagreement. For suppose every disagreement were faulty, i.e., suppose that in every case of disagreement, at least one party must be wrong. If, as seems natural, we identify making an assertion that is wrong with asserting a (simply) false proposition, then if such were the case relativism (moderate or not) would be unmotivated: we would not need a notion such as relative truth to make sense of exchanges such as (1) to (2). Simple truth and falsity would do. If on the other hand every case of disagreement were faultless, then radical, not moderate relativism would be motivated as the best explanation: For in such a case (under the same assumptions) no propositional content could ever be regarded as simply false. Again, if apparent cases of faultless disagreement were not cases of disagreement (i.e., both parties' assertions can be plainly true) then relativism—moderate or radical—would be unmotivated. Moderate relativism is motivated only if:

- there are genuine cases of faultless disagreement, and
- not every case of disagreement is faultless.

Moderate relativists explicate the "sense" in which Ann and Bill are both right by claiming that both of their assertions are true, though relatively so: It is true *for Ann* that Korean cabbage is delicious, while it is true *for Bill* that it is disgusting (hence, not delicious). Both Ann and Bill are right in the sense that the content Ann asserts is true (for her), and the incompatible content that Bill asserts is also true (for him). However, the notion of relative truth, or truth-for, that is thereby introduced is not without problems. Not that we do not often say such things as "It is true for Ann that Korean cabbage is delicious," or, perhaps even more frequently, "For Ann, it is true that Korean cabbage is delicious." The issue is, first, what we mean by such statements, and, secondly, on the hypothesis that (in some cases at least) what we mean involves the notion of truth-for, what truth-for *is* exactly and how does it relate to plain, monadic truth. John McFarlane put the latter issue very well in one of his seminal chapters on relative truth:

> …it is not clear that the concept of truth *admits* of relativisation to assessors. If 'true' as it occurs in 'true for *X*' is just the ordinary, non-relative truth predicate, then it is unclear what 'for *X*' adds, unless it is just 'and *X* believes this'. On the other hand, if the occurrence of 'true' in 'true for *X*' is like the 'cat' in 'cattle', then the relativist needs to explain what 'true for *X*' means and what it has to do with truth, as ordinarily conceived. (2005, p. 328)[2]

In this chapter, I will take up three side issues that are, however, of some interest in the way of bush-beating. They are (1) whether relative truth can be explicated in terms of subjective worlds, (2) whether alleged relative truth could just be belief (i.e., "*p* is true for *X*" = "*X* believes that *p*"), and, finally (3) whether plain truth could, or should be defined on the basis of relative truth.

10.2 First Point: Relative Truth as Truth in Subjective Worlds

We have no trouble understanding what it is for a *sentence* to be true or false at some value of some parameter. For example, it is not hard to understand what is meant by saying that the sentence "I am lazy" may be true at one index and false at another. It may be explicated by pointing out that *different facts* make the sentence true depending on the context in which it is used—more particularly, depending on who is uttering it. If one does not like facts, it may also be explicated by pointing out that *different inferences* may be sustained by different uses of that same sentence. The basic, relatively clear intuition is that the content I express by uttering "I am lazy" is different from the content you express by uttering "I am lazy."

With relativity of a *proposition's* truth value to an extra parameter (judge, point of evaluation, etc.) things are not equally straightforward. For example, it seems that here we could not say that the proposition sustains different inferences, depending on the parameter's value: The inferential potential of *Korean cabbage is delicious* appears to be the same no matter who is evaluating it (which is as it should be, for inferential potential depends on content, but content is supposed to be unaffected by the extra parameter). A contextualist might object that the inferential potential of "Korean cabbage is delicious$_{\text{Ann}}$" is not the same as the potential of "Korean cabbage is delicious$_{\text{Bill}}$" as (e.g.,) the former but not the latter implies that Ann likes Korean cabbage. However, the proposition(s) whose potential the contextualist is assessing is not the same as the one the moderate relativist is considering; variation in inferential potential does not depend on who *evaluates* the proposition.

Could we say that the proposition is made true by different facts? Crispin Wright has considered the hypothesis that we could:

> We might for example permit the actual world *at Williamson* —that is the actual world as reflected in Timothy Williamson's gustatory standards—to exist simultaneously with the actual world *at Wright*, that is, the actual world as reflected in Crispin Wright's gustatory

[2] For another statement of the prima facie implausibility of the very notion of relative truth see Richard (2004, p. 226, p. 230).

> standards. The proposition that stewed rhubarb is delicious can then be true at the one aesthetic location, so to say, and untrue at the other […]. No need, in this case, to resort to a context of assessment parameter in order to accommodate the truth-relativist impulse—truth can be old-fashioned truth-at-a-world, *simpliciter*. The relativism surfaces, rather, in the thought that there is no single actual world but a plurality of them. (2008, p. 172)

The idea is that we might countenance *subjective worlds*: A subjective world is the actual world as distorted by someone's standards—standards of taste, in our case, or, in other cases, aesthetic or epistemic standards. The same proposition will be said to be true—plainly true, not relatively true—at one subjective world but not at another. This, it is suggested, is one way we could make sense of relative truth: for a content to be true *at X* is for it to be (simply) true at some *X*-distorted world. As with the relativity of a sentence's truth value, different facts make one and the same content true, depending on whose standards we are considering: except that here the relevant facts belong to different subjective worlds.

Several objections can be raised against this view as a way of making sense of relative truth. Let us start with an objection that, though inaccurately put, does point to a genuine difficulty. Annalisa Coliva and Sebastiano Moruzzi (2010, unpublished) have argued that on the "subjective worlds" view alleged cases of faultless disagreement would not be cases of disagreement; hence, the view must be unacceptable to the moderate relativist. They say:

> If *P* and not-*P* are true in different worlds, in what way could their respective supporters be in disagreement with each other? For, in order to have disagreement…they should maintain that they are both true in the same world.

Now, that two assertions *p*, *q* are true at different worlds surely does not entail that the asserters cannot be in disagreement with each other. Suppose Williamson asserts that Aconcagua is more than 7,000 m tall, while Wright asserts that Aconcagua is not more than 7,000 m tall. Their assertions are made true by different worlds (Williamson's by the actual world among others, Wright's by some other possible world), yet they obviously disagree with each other. The reason we say they disagree is that we take their assertions to be about the same world, i.e., the actual world, and they cannot both be true in the same world. We would say they do not disagree if their assertions were *taken to be about* different worlds, not just true in different worlds.

One way we would take their assertions to be about different worlds would be by reading them as containing an implicit "subjectivity" operator: For example, we might take Williamson's assertion to amount to "*In my world,* stewed rhubarb is delicious," and similarly for Wright's assertion (call this "the subjective reading" of assertions of taste). Such a reading would be a variety of contextualist reading: It would introduce a contextual parameter, saturation of which generates compatible contents. Contents would be compatible in that they can both be true in the same world, as is usually the case with contextualist readings: For example, that Korean cabbage is delicious *according to Ann's taste* is compatible with its being disgusting *according to Bill's taste,* as both may be actually the case. In our case, assuming that it is true in the actual world that in *X's* world stewed rhubarb is delicious if and

only if it is (simply) true, in *X's* subjective world, that stewed rhubarb is delicious,[3] it would follow that the content of Williamson's assertion is compatible with the content of Wright's assertion, as both can be true in the actual world. As Coliva and Moruzzi noticed, disagreement evaporates.

Obviously, a moderate relativist would reject the subjective reading (as she would reject all contextualist readings of assertions of taste) exactly because it makes disagreement disappear. According to the moderate relativist, this is enough to show that the subjective reading misinterprets Williamson's and Wright's assertions. Is there some other way in which the moderate relativist can avail himself of Wright's subjective worlds? Surely he could not say that the assertion that stewed rhubarb is delicious is true for Williamson (or "at" Williamson) if and only if Williamson's subjective world w_T *coincides* with the actual world w^* (at least) in the fact that stewed rhubarb is delicious, i.e., that w^* belongs to the existence set of "Stewed rhubarb is delicious" (Mulligan and Correia 2007, p. 9). For that would amount to the actual existence of the fact that stewed rhubarb is delicious: So on the one hand, the appeal to subjective worlds would be redundant; on the other, the relativist would be retracting his constitutive claim that "there are no facts of the matter" as to whether, say, stewed rhubarb is delicious (Kölbel 2002, p. 19; Lasersohn 2005, p. 644; Wright 2006, p. 52). For the (allegedly relative) truth of the assertion that stewed rhubarb is delicious would entail that it is an actual fact that it is. Any disagreement with the assertion would then be bound to be faulty.

A more promising way for the relativist to avail himself of subjective worlds would start by saying that though assertions of taste are *not* intended to be about subjective worlds, they are to be *evaluated* at subjective worlds: Williamson's assertion that stewed rhubarb is delicious is true (for Williamson) iff it is simply true at w_T. One could then choose between leaving the notion of simple truth at w^* undefined for such assertions, and defining a notion of *subjective* truth at w^* to the effect that p is subjectively true (at w^*) iff there is a subjective distortion of w^*, w_X, such that p is simply true at w_X. Assertions of taste, among others, could never be simply true at w^*, only subjectively true. It seems to me that this would be the use of subjective worlds that best correponds to Wright's suggestion. To be sure, it was part of the suggestion that "there is no single actual world": if the relativist wants to stick to it, he will choose the first option and leave simple truth at w^* undefined for assertions of taste (etc.). Alternatively, the relativist might pick up another part of Wright's suggestion (Williamson's subjective world "is *the actual world* as reflected in Timothy Williamson's gustatory standards," my it) and grant that there is an actual world, though no assertions of taste are simply true at it—only subjectively true, that is, true in some subjective distortion of it. This framework may be seen as accounting for disagreement, in that there is no world at which both contents (Williamson's and Wright's) are simply true. Neither content is simply true at the actual world, as on both options simple truth at the actual world is not defined for contents of that kind; moreover, they cannot be both simply true at any subjective world, if subjective worlds are assumed to be consistent. If "stewed rhubarb is delicious" is true at Williamson's world w_T then "stewed rhubarb is disgusting" is not true at w_T.

[3] Reading the implicit operator "in my world" as both modal and indexical.

However, even the more promising framework is not without difficulties. First of all, it can be argued that it does not seriously account for disagreement. Plausibly, Williamson and Wright disagree about whether stewed rhubarb is delicious or disgusting. Now, in the actual world stewed rhubarb does not possess the property of being delicious: If it did, Williamson would be right and his disagreement with Wright would not be faultless. Stewed rhubarb is delicious (or disgusting) only in subjective worlds. Hence, there is nothing in the actual world for Williamson and Wright to disagree about; more precisely, their disagreement is not about whether a property that could be instantiated by stewed rhubarb is actually instantiated by it. Yet, this is exactly what their disagreement appears to be about. That the contents they express by their assertions are incompatible (i.e., there is no world at which they are both true) does not illuminate the nature of their disagreement.

The second difficulty concerns the inherent plausibility of subjective worlds as ontological constructions. On most views, worlds consist of facts (Mulligan and Correia 2007, p. 9): if there are subjective worlds, then there are subjective facts. Facts, in turn, are usually regarded as *correctness* conditions for judgments, assertions, and beliefs:

> if *x* judges correctly that *p*, then the state of affairs that *p* obtains. And if this is plausible, so too is the further claim that if x judges correctly that *p*, then *x* judges correctly that *p* because the state of affairs that *p* obtains. Facts make judgments correct. (Mulligan and Correia 2007, p. 6)

Thus the subjective fact (of w_T) that stewed rhubarb is delicious makes Williamson's assertion correct. Now, suppose this is the case for each and every assertion of taste that Williamson makes: No matter what assertion Williamson makes in matters of taste, there is a fact in w_T that makes it correct. If such were the case, Williamson's subjective world w_T would be like a mirror of his assertions of taste. But then, the claim that facts are correctness *conditions* would be in question: For it seems to be part of the very idea of a correctness condition that it may fail to hold, thereby making an assertion incorrect. This is part of the intuitive idea of the *resistance* of facts: facts are there to possibly belie our judgments, assertions, and beliefs. If subjective facts do not possess this capacity then they are funny facts indeed (and subjective worlds are funny worlds).

But perhaps we do not need to assume that subjective worlds are mirrors of a subject's assertions of taste. Among relativists, Max Kölbel (2008) has claimed that it is possible for someone to have false beliefs about her tastes. For example, Ann may believe that whale meat is not tasty, whereas in fact whale meat is tasty for her (perhaps she never actually tried whale meat but, somehow, she came to acquire the belief that it is disgusting). If Kölbel is right, Ann might sincerely assert that whale meat is not tasty while it is a fact of her subjective world that it *is* tasty. On this assumption, subjective facts would genuinely play the role of correctness conditions for assertions of taste.

As we shall see below (§ 3), Kölbel's suggestion may be based on a naturalistic (mis)interpretation of what it is for something to be tasty (disgusting, funny…) *for* a subject: An interpretation that a relativist should really reject, for it has the consequence of eliminating disagreement in matters of taste (see below). But suppose

for a moment that the hypothesis is relativistically acceptable. Subjective facts do not simply mirror assertions of taste: one can be wrong about one's tastes. That Williamson sincerely asserts that stewed rhubarb is delicious does not guarantee that his assertion is true in w_T, as w_T may or may not contain the fact that stewed rhubarb is delicious, depending on what Williamson's tastes *really* are: w_T might well contain the fact that stewed rhubarb is disgusting. Even so, there is a difficulty.

Let us reflect on what determines the facts that constitute a subject *S's* subjective world w_S. w_S is supposed to reflect *S's* standards of taste: It is a fact that p in w_S if and only if, on *S's* standards, p. It is essential to realize that its being the case that p according to *S's* standards is both necessary and sufficient for its being the case that p in *S's* subjective world w_S.

This being so, it seems that, once more, subjective facts are somewhat *sui generis*. For on the one hand, as p is a fact (albeit subjective), we have:

(3) If it is correct (on *S's* standards) that p, then it is correct that p because it is a fact (of w_S) that p.

But on the other hand:

(4) If it is a fact (of w_S) that p, then it is a fact (of w_S) that p because it is correct (on *S's* standards) that p.

One may of course have legitimate doubts about both occurrences of "because," in (3) as well as in (4). However, the occurrence in (3) is motivated by minimal truth making, "the widespread intuition that truth [as well as correctness] is truth *in virtue of* something" (Mulligan 2006, p. 34). As to the occurrence in (4), why should we assume that it is a fact of Williamson's subjective world w_T that stewed rhubarb is delicious, were it not for the fact that stewed rhubarb *is* delicious according to his standards? It may be objected that this is merely our reason for assuming that such a fact exists in w_T, not the cause of its existence. However, in both (3) and (4) "because" may be explanatory without being causal.

Now, that (3) and (4) both hold will not do: "instances of 'p because p' are all false" (Mulligan 2006, p. 32); particularly, explanations cannot go both ways. This difficulty expresses what can be seen as the essential problem with subjective facts: On the one hand, they are there to make assertions of taste true (though subjectively so); but on the other hand, they are a mere reflection of a subject's standards and judgments of taste. It seems unlikely that one entity can do both jobs.

10.3 Second Point: Relative Truth and Belief

When trying to make sense of the notion of relative truth, or truth-for, one easily falls prey to the temptation of taking relative truth to be just *belief*: That is, saying that a proposition p is true for X amounts to saying that X believes that p. One rather obvious reason for yielding to the temptation is the following. At least *prima facie,* moderate relativists seem to regard a subject *A's* assertion that p as sufficient ground for the semantic statement "p is true for A." Clearly, this sets relative truth apart

from plain truth, as we do not regard a subject's sincere assertion that *p* as sufficient ground for the semantic statement "*p* is true" (it is a fact of life that asserters are often mistaken). On the other hand, a subject's sincere assertion that *p* is usually regarded as sufficient ground for attributing her the belief that *p*. In this respect, and in spite of superficial appearance, "*p* is true for *X*" turns out to be closer to "*X* believes that *p*" than to "*p* is true."

Aside from that, the belief paraphrase seems to capture certain common, though admittedly vague relativistic intuitions: After all, many instinctive relativists appear to think that in some domains of discourse such as judgments of taste and aesthetic judgments there are no genuine truths, only opinions.[4] With some effort, the intuition could be extended to knowledge ascriptions ("It is true for *X* that *Y* knows that *p*" = "Given *X's* beliefs, *Y* knows that *p*").

However, at least some moderate relativists have no use for such a paraphrase. Among them, Max Kölbel has been particularly explicit and persistent in rejecting it (Kölbel 2002, pp. 33–34, 2008, p. 19, 2009, pp. 393–394). His argument is that a proposition *p* might be true for a subject *X* even though her beliefs are incompatible with *p,* or *X* has no beliefs at all concerning *p*. For example:

> It seems entirely coherent to say [WM] 'John has no view as to whether whale meat is tasty. But in fact whale meat is tasty for him.' (2009, pp. 393–394)

But on the belief interpretation of "true for," (WM) would be incoherent. Or again:

> People…sometimes believe or assert propositions that are not true according to their own standard. For example, Barbara may come to believe, as a result of listening to Anna's utterance ["Depp is more handsome than Pitt"], that Depp is more handsome than Pitt. She might later realize that this belief is a mistake, because she prefers Pitt. (Kölbel 2008, p. 19, fn. 30)

We may suppose that was already the case at the time *t* at which Barbara believed that Depp was more handsome than Pitt: That is, at *t* she believed that Depp was more handsome than Pitt, while it was true for her (at *t*) that Pitt was more handsome than Depp. On the belief paraphrase of "true for," she would have entertained inconsistent beliefs.

Now, there is little doubt that we do say such things as "John has no view as to whether whale meat is tasty. But in fact whale meat is tasty for him" without perceiving any inconsistency. But the reason may be one the moderate relativist would not approve of: We might be interpreting "Whale meat is tasty for him" on the pattern of—say—"Gluten is toxic for her," i.e., as asserting that a certain natural relation, *being tasty for,* holds between whale meat and John. Now, perhaps "tasty (for *x*)" is occasionally taken to stand for some such natural, wholly objective relation: On this reading, that something is tasty for someone is entirely objective, as objective as a substance's toxicity for a given organism. However, if "tasty" had this kind of semantics there would be no room for a relativistic account of "*x* is tasty" If "tasty" stood for a natural relation between a substance (such as whale meat) and a person, then the assertion that whale meat is tasty would necessarily be elliptic

[4] Such vague intuitions are reported by Lasersohn (2005, p. 643) among others.

for the assertion that whale meat is tasty for ( ): There would be no content "whale meat is tasty" to be evaluated by a judge or from a perspective. Moreover, the (completed) assertion that whale meat is tasty for John would state that whale meat has a certain objective effect on John, not that whale meat has the property of tastiness as evaluated from John's perspective, or taking John as the judge, etc. Similarly, exchanges such as (1) to (2) would not lend themselves to a relativistic treatment. They would involve no disagreement, as both (1) and (2) ought to be read contextually and could be both true (or both false, as the case may be).

I am not claiming that examples such as (WM) *must* be given the "naturalistic" reading if they are to be regarded as consistent; there may be other possibilities. However, the suspicion is strong that the consistent reading of (WM) may depend on a naturalistic reading of "tasty." Hence, a moderate relativist had better not rest his case against the belief paraphrase on this kind of examples.

Anyway, there is a stronger argument against the paraphrase. As John MacFarlane remarked some time ago (2005, p. 328), there are constraints on what "true for" can be taken to mean within the framework of moderate relativism. Let us be reminded that moderate relativists countenance *both* simply true propositions *and* propositions that can only be said to be true *for* some point of view (judge, etc.), though not for some other. Given this assumption, could relative truth be a subspecies of simple truth, that is, could "*p* is true for *X*" mean "*p* is (simply) true and *R*(*X*, *p*)?" It seems not; for in such a case, every relative truth would be a simple truth as well, so that it's *not* being a truth for someone, *Y,* could only be brought back to some deficiency on *Y's* side. This would make *Y's* disagreement faulty rather than faultless, thereby voiding the moderate relativist's motivation. If on the other hand *every* relative truth is universally relative (if *p* is true for *X*, there is no *Y* such that *p* is not true for *Y*), then again there are no instances of faultless *disagreement* (as there are no cases of disagreement): the same conclusion follows.

Call "Autonomy" the requirement that relative truth is not a subspecies of truth. According to Autonomy, not *every* relative truth is a simple truth as well. Could it be that *some* are, that is, could there be a relative truth *p* that is also a simple truth, or must the extensions of relative truth and simple truth be disjoint? We shall later see that some relativists countenance the former possibility. However, I believe they should accept a weaker requirement: There cannot be propositions that are *both* simply true *and* "essentially controversial", that is, true relatively to *X* but not true relatively to *Y* (for some *X* and *Y*). If such were the case for some proposition *p,* then, as *p* is simply true, its not being true for Y could only derive from some inadequacy of *Y,* as in the previous case: *faultless* disagreement would not arise (hence *p* could not be true relatively to *X,* contrary to the hypothesis). Call this "the (Weak) Disjoint Extensions requirement." Autonomy does not entail WDE, as on Autonomy there might be some propositions that are both simply true, true for *X,* and not true for *Y*. Strictly, neither does WDE entail Autonomy. Suppose Autonomy does not hold, so that "*p* is true for *X*" amounts to "*p* is simply true and *Q*(*p*)" (for some *Q*), hence, every relatively true proposition is simply true as well; nonetheless, it could still be the case that no such proposition is also not true for *Y* (for some *Y*), namely, if no relatively true proposition were controversial. In that case, however,

relativism would be unmotivated. So we could say that WDE entails Autonomy (and is stronger than Autonomy) on condition that at least one relatively true proposition is controversial.

Now, let us consider again the belief paraphrase, "*p* is true for *X*" = "*X* believes that *p*." Clearly it does not violate Autonomy, as it does not entail that every relatively true proposition is simply true as well. But, on the other hand, it violates WDE, for the following reason. The moderate relativist admits that there are propositions that are simply true. Let *p* be one such proposition, for example, a tautology or an arithmetical truth. It is plausible to assume that for any such truth *p,* there may be someone who does not believe that *p*. Maybe she does not understand it, or perhaps she just fails to see the point. On the other hand, there will be lots of other people who do believe that *p*. Hence, on the belief paraphrase, for any such simply true *p, p* will be true relatively to *X* (for many *X*) and not true relatively to *Y*: WDE is violated. This finally condemns the belief paraphrase.

But if relative truth cannot be understood in terms of belief, how can we otherwise motivate the relativist's basing the semantic statement "It is true for Ann that Korean cabbage is delicious" on Ann's presumably sincere assertion "Korean cabbage is delicious?" In fact, we do not need any such motivation, for the relativist's semantic judgments can be seen in a different light. The relativist need not conceive of semantic statements such as (5)

(5) It is true for Ann that Korean cabbage is delicious

as entailed by sincere assertions; he may see them—and relativist semantics in general—as providing *the best explanation* of certain intuitive features of exchanges like (1) to (2), namely of their being cases of faultless disagreement.[5] If we assume that judgments of taste have a relativistic semantics, we can see how Ann and Bill can disagree without either being at fault. Ann's sincere assertion (1) may in itself provide only a *prima facie* reason for the semantic statement (5): Depending on how relative truth is conceived, it may well turn out that, though Ann sincerely asserts that Korean cabbage is delicious and believes as much, it is *not* true *for* her that it is. In other words, that Ann's assertion of (1) is a *prima facie* ground for (5) does not commit the relativist to a notion of relative truth on which Ann's assertion of (1) automatically makes (5) true.

10.4 Third Point: Simple Truth as Universally Relative Truth

WDE is the requirement that for no proposition *p, p* is both simply true *and* true relative to *X* (for some *X*) and not true relative to *Y* (for some *Y*). Should the relativist also accept a stronger constraint, that is, that for no *p, p* is both simply true *and*

[5] This is in agreement with Crispin Wright's (2006, p. 42) suggestion that relativism is best seen as a theoretical attempt to make sense of the properties (= faultless disagreement) that exchanges such as (1) to (2) are ordinarily attributed.

true relative to *X* (for some *X*)? Call this the "Strong Disjoint Extensions" requirement (SDE). Considering the moderate relativist's general attitude, SDE makes sense: for his view seems to be that there are propositions, such as those expressed by judgments of taste etc., to which the notion of simple truth just does not apply (they are not the kind of contents that *could* be simply true or simply false), whereas other propositions – "7 + 5 = 13", say—are just true or, as in this case, just false. As I already pointed out, most contemporary relativists seem to agree that relativistic semantics is in order for some contents though not for others: "the relativist views in contemporary debate are typically local" (Wright 2008, pp. 167–168). Thus Lasersohn restricts relativism to sentences involving "subjective" predicates, i.e., predicates that generate disagreement that cannot be decided by objective facts (2005, pp. 682–683); and Richard takes relativism to apply to expressions that are subject to "accommodation and negotiation" (2004, p. 228). As relativist semantics only applies to certain linguistically individuated contents, it follows that other contents are regarded as standard in point of truth, that is, simply true or simply false. Within this frame of mind, a proposition is *either* the kind of content that is simply true or simply false, *or* the kind of content that can only be true *for* someone though possibly not *for* someone else. No proposition can belong to both kinds. Simple truth is not conceived as "truth in every perspective," that is, as universally relative truth: The notion of relative truth just does not apply to ordinary, nonsubjective contents.

With respect to this consensus, Max Kölbel stands apart: he takes relative truth as basic and defines a notion of *objective* truth on the basis of relative truth (objective truth is "necessarily universal" relative truth). It follows that the notion of relative truth does apply to contents that are objectively true. Kölbel first defines objectivity (2002, p. 102):

> (OBJ) For all *p*: *p* is objective iff it is not possible that there be thinkers *A* and *B,* such that *p* is true in *A's* perspective and *p* is not true in *B's* perspective.

A proposition is then said to be *objectively true* iff it is both objective and true for some *X* (hence for every *X*); it is objectively false iff it is both objective and false for some (hence for every) *X*. Thus, objectively true propositions are true for some *X*—that is, relatively true—as well.

Is objective truth a reasonable *explicatum* of simple truth?[6] It does play one role of simple truth: If a proposition is objective (objectively true or objectively false), then any disagreement on it is bound to be faulty. Moreover, there is some *prima facie* plausibility in the view that simple truth is just truth *for* everyone. However, Kölbel's objectively true propositions are not just relatively true propositions that *happen* to be true for everyone. There is no inconsistency in holding that all perspectives might happen to agree on some content, though it is a content inherently suitable for perspectival truth: For example, it might so happen that Barolo is an excellent wine in all perspectives. If so, Barolo would be universally perspectivally excellent. However, Kölbel's objective truth is not *accidentally* universal relative truth: An objectively true proposition is objective, that is, it is such that it is a priori

[6] In Kölbel's (2009) terminology it clearly is not, as he reserves the phrase 'simple truth' for the *relative* notion.

that, if it is true for some *X,* then it is true for every *X*. That "it is not possible that there be thinkers *A* and *B,* such that *p* is true in *A's* perspective and *p* is not true in *B's* perspective," Kölbel explains, means that "it is ruled out by a priori constraints on language use" that *p* might be true for *A* but not for *B* (2002, p. 139, fn 15). In other words, objective contents are singled out by the linguistic features of sentences that express them. Here, Kölbel seems to agree with the consensus: Contents to which simple truth and simple falsity apply are distinct from "inherently controversial" contents—contents that might be true for *X* but not for *Y*. But then—one wonders—what is the point of applying the notion of relative truth to contents that could not possibly be true for one but not for another? The notion of accidentally universal relative truth does make sense: any number of debaters may turn out to agree on an inherently controversial content. Barolo may turn out to be good for everyone (though, in principle, someone might disagree). However, it is not equally clear that the notion of *necessarily* universal relative truth makes sense. If *p* is inherently incapable of generating faultless disagreement because of the kind of content it is (as determined by a priori constraints on language use), why should we want to say that, nevertheless, it is true *for X* (or false for *X*) rather than simply true, or false? In moderate relativism of all varieties (including Kölbel's), relativist semantics is motivated by faultless disagreement: where faultless disagreement is a priori ruled out, application of the notion of relative truth is unmotivated.

Acknowledgments Thanks to Andrea Iacona for useful criticism of a previous version.

References

Boghossian P (2006) Fear of knowledge. Oxford University, Oxford
Cappelen H, Hawthorne J (2009) Relativism and monadic truth. Oxford University, Oxford
Coliva A, Moruzzi J (2010) Is there a coherent notion of relativism? Unpublished paper
Kölbel M (2002) Truth without objectivity. Routledge, London
Kölbel M (2008) Introduction: motivations for relativism. In: Garcia CM, Kölbel M (eds) Relative truth. Oxford University, Oxford, p1–38
Kölbel M (2009) The evidence for relativism. Synthese 166:375–395
Lasersohn P (2005) Context dependence, disagreement and predicates of personal taste. Linguist Philos 28:643–686
Lopez de Sà D (2008) Presuppositions of commonality: an indexical relativist account of disagreement. In: Garcia Carpintero M, Kölbel M (eds) Relative truth. Oxford University, Oxford, pp 297–310
MacFarlane J (2005) Making sense of relative truth. Proc Aristot Soc 105:321–339
Mulligan K (2006) Facts, formal objects and ontology. In: Bottani A, Davies R (eds) Modes of existence. Ontos, Frankfurt, pp 31–46
Mulligan K, Correia F (2007) Facts. The Stanford encyclopedia of philosophy. Available with: http://plato.stanford.edu/entries/facts/). Accessed 29 May 2012
Recanati F (2007) Perspectival thought: a plea for moderate relativism. Oxford University, Oxford
Richard M (2004) Contextualism and relativism. Philos Stud 119:215–242
Wright C (2006) Intuitionism, realism, relativism and rhubarb. In: Greenough P, Lynch MP (eds) Truth and realism. Oxford University, Oxford, pp 38–60
Wright C (2008) Relativism about truth itself: haphazard thoughts about the very idea. In: Garcia CM, Kölbel M (eds) Relative truth. Oxford University, Oxford, pp 157–185

Chapter 11
Mere Belief as a Modification

Maria van der Schaar

Abstract The chapter uses the method of linguistic phenomenology to explain how belief in the sense of mere opinion can be understood as botched knowing. The distinction between attributive and nonattributive terms plays a central role in this explanation of belief. Several kinds of nonattributive terms are distinguished, modifying, restrictive and restorative terms, each being of use in the explanation of epistemic notions. And several forms of modification are distinguished: semantic, conceptual and ontological modification.

Keywords Belief · Knowledge · Modification · Linguistic phenomenology

11.1 Introduction[1]

My *Doktorvater* Gabriel Nuchelmans was an advocate of analytic philosophy in the Netherlands, but his innovative research was primarily related to the history of philosophy. Writing my thesis, I was in need of someone who was doing original work in philosophy, who shared my interest in analytic philosophy and phenomenology and who understood that philosophy and its history are in need of each other. At a summer school on Austrian philosophy in Bolzano, in 1988, I told Kevin I was working on a thesis on G. F. Stout, and we stayed in touch. I could spend some time in Geneva living with his students, who were renting a house near the border in France. Notwithstanding a wide variety of topics we were working on, we were all engaged in doing linguistic phenomenology, as Austin puts it in his paper 'A Plea for Excuses' (Austin 1956, p. 182). We learned that analytic philosophy is in need

[1] This is a second version of the paper, originally called 'Mere Belief and the Etiolations of Language', which was published in the online Festschrift for Kevin. The section on the etiolations of language in the paper has been rewritten and expanded into an independent paper, published as van der Schaar (2014). I therefore decided to rewrite that section completely, only summarizing the results of the earlier section. Because there was now some space left, I could rewrite the 'Mere belief' paper, and because Kevin had given his comments on that paper to me, I have taken the opportunity to elaborate on some ideas in the original paper.

M. van der Schaar (✉)
Institute for Philosophy, Leiden University, Leiden, Netherlands
e-mail: m.v.d.schaar@phil.leidenuniv.nl

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_11,

of more examples than ‘The morning star is identical with the evening star’, ‘The king of France is bald’ and ‘A bachelor is an unmarried man’, and that many of the fruitful insights in analytic philosophy were predated in phenomenology. Kevin’s sensitivity to the varieties of language and experience was an example to us.

Although the explanation of knowledge in terms of justification, truth and belief has been criticized since Gettier, the criticism has not been directed at the explanation of knowledge in terms of belief. Timothy Williamson is an exception, for he takes knowledge to be a primitive notion, and explains mere belief as a kind of botched knowing (Williamson 2000, p. 47). The term ‘botched’ is a modifying term, like the term ‘fake’ or ‘false’: A false Rembrandt looks like a Rembrandt, pretends to be one, but is not a Rembrandt. In standard cases, adjectives are attributive: The term ‘red’ in ‘red jacket’ is attributive, because it is used to attribute the quality of being red to the jacket, in such a way that a red jacket is a special kind of jacket. Modifying terms are nonattributive, because these terms are not used to attribute a quality to the object denoted by the noun: We do not attribute the property of being false to the painting, although we do attribute to it the property of being a false Rembrandt, when we claim that it is. Other kinds of nonattributive terms are ‘mere’, ‘true’, ‘actual’ and ‘real’. Neither a false nor a true Rembrandt is a special kind of Rembrandt, but the latter is a Rembrandt nonetheless. Nonattributive terms play an important role in phenomenology in the explanation of knowledge, judgement and intentionality. Interest in the etiolations of language can also be found in J. L. Austin’s *How to Do Things with Words, Sense and Sensibilia* (1964) and ‘Other Minds’ (1946). How can one use the phenomenological method and the linguistic method developed by Austin to elucidate the relation between knowledge and belief, especially the idea that mere belief is a form of botched knowing?

11.2 Four Ways to Relate Knowledge and Belief

(i) Knowledge may be explained in terms of belief. If someone says: ‘I know that dogs descend from wolves, but I don’t believe it’, we rightly call him irrational. And we may react: ‘If you know it, how is it possible that you don’t believe it?’ Because belief is a necessary condition for knowledge, one might be tempted to understand knowledge as a special kind of belief, and explain knowledge in terms of belief. There are, though, some problems with this order of explanation. A general rule of defining is that the less clear notion should be explained in terms of the clearer notion. It seems, though, that the term ‘belief’ is not at all clearer than the term ‘knowledge’, for ‘belief’ has several meanings, which all seem to be relevant to the notion of knowledge.

The term ‘belief (that)’ may mean:

a. (A capacity to) judge, which is an all or nothing affair[2]
b. Conviction, which has degrees

[2] The term ‘judgement’ itself may stand for the act of judgement, the judgement product, the judgement candidate or the faculty of judgement; cf. van der Schaar (2007).

c. Opinion, which is opposed to knowledge
d. Unquestioned faith, a kind of trust[3]

If knowledge is explained in terms of belief, the meaning of 'belief' as opinion is excluded. If knowledge is explained in terms of justification or a related notion, the meaning of 'belief' as unquestioned faith seems to be excluded, too. So, we need to focus on meaning (a) and (b). It seems that both meanings play a role in the explanation of knowledge: A necessary condition for knowing that *S* is that we judge that *S,* or that we have once judged that *S,* and that this judgement is stored in our memory in such a way that we are able to make the judgement again. Furthermore, a minimal degree of conviction is also a necessary condition for knowledge. In modern analytic philosophy, knowledge and belief are generally understood as states of the mind, but what a state of mind is, is not explained. 'Belief' in sense (a) and 'belief' in sense (b) are sometimes called 'mental states', but not in the same sense. Belief in the sense of conviction can be called a state of mind, in the way we call doubt a state of mind (cf. Reinach 1911, p. 320). Such a mental state extends over a certain period of time, and its temporal parts are homogenous, that is, these parts are of the same quality. A mental state of being convinced that *S* may have different degrees, and may be called a feeling. A feeling of certainty is a high degree of confidence. All our judgements are accompanied by a certain degree of conviction, that is, a certain feeling. Such a feeling disappears as soon as we fall asleep. Because a state of mind extends over a certain period in time and has homogenous parts, a mental act should be distinguished from a mental state: acts are not extended in time the way states are. An act of judgement, for example, is not a mental state. As a silent act of assertion, it belongs to the same mental category as the speech act of assertion. Some acts seem to be stretched out in time, such as an act of proving, but in these cases the parts of the act are not homogenous. Furthermore, only the final moment makes the act an act of proving. If such a final moment is not obtained, the act is merely an act in which one purported to prove something. An act, it is true, but not one of the right kind.

Another point of difference between acts and states is that acts are internally related to their products: The act of proving results in a (proven) theorem, and the act of building a house results in the house built. Equally, an act of writing a letter results in a written letter, an act of promising in a promise made, an act of assertion results in an assertion made and an act of judgement results in a judgement made (cf. Twardowski 1912). States such as a state of doubt are not thus internally related to products. Being in a state of doubt has a beginning and an end, which are homogenous to the other temporal parts of the state.

Philosophers say that a belief in sense (a) is a state of mind. A belief, they say, has a physiological counterpart, a disposition, which causes us to have certain thoughts. Such a disposition they call a *state,* and its mental counterpart is called a *mental state*. Such a naturalist approach to belief is not able though to explain in what sense beliefs can be called right or wrong; for, as dispositions we can at most say that they

[3] These different meanings of the term 'belief' are given an account of in van der Schaar (2009). Unquestioned faith is perhaps not a case of belief that, but of belief in. Propositional truth and falsity do not seem to be applicable to the content of unquestioned faith.

exist, or that they do not exist (cf. Hacker 2004). This way of looking at judgement makes it difficult to give an explanation of the capacity to judge. For, it seems to be an unconscious cause of our acts of judgements, on this account, and it seems that we have to understand such a cause, before we can understand what the act of judgement is. Neither the act of judgement, as we have just seen, nor the capacity to judge can be called a mental state, and the terminology is therefore inapt.

Capacities should be understood in terms of their actualizations, rather than in dispositional, causal terms. It belongs to the essence of a capacity, being a potentiality, that it can be actualized, and the explanation of a capacity is therefore to be given in terms of its actualizations. One has to make a distinction between a general and a specific capacity. The boy who is able to make more complex calculations has the general capacity to come to know the sum of 67 and 88, and has in this sense the capacity to judge, to come to know, that the sum of 67 and 88 is 155. As long as he has not made the calculation, though, he does not have the specific capacity to judge that 67 and 88 is 155. Only the specific capacity to make the judgement is standardly called a 'belief', where the term is to be taken in sense (a). The belief that *S* in the sense of the capacity to judge that *S* can now be explained as: One has once judged that *S,* and one has restored this judgement in one's memory in such a way that one judges that *S* in appropriate circumstances. And knowledge can be understood as a special case of belief in this sense. In another sense of 'to know', the boy may be said to know the sum of 67 and 88, as soon as he has mastered the general capacity. He knows *how* to make this kind of calculations.

Those philosophers who call an act of judgement an 'occurrent belief' may understand the relation between belief and act of judgement in two ways: either they consider belief to be a general term covering both capacities and acts, or they consider 'belief' to cover primarily a capacity, or a disposition, and 'occurrent belief' to be an actualization or a manifestation of this capacity or disposition. On the latter account, an occurrent belief is not a special case of belief, but an expression of it, just as coughing may be the expression of a cold. On either account, the act of judging is explained in terms of the dispositional notion belief, or the capacity. There is a reason, though, to prefer the Aristotelian order of explanation, in which belief is understood as a capacity or potentiality, and, because a potentiality is a potentiality to be actualized, the potentiality is to be explained in terms of its actualization. Instead of calling an act of judgement an occurrent belief, it is thus preferable to call a belief a capacity to judge, and the *act of judgement* is thus prior in the order of explanation to *belief*.

The verb 'to know' can be used for an act of knowing that results in a certain product, and for knowledge as capacity, as well. Acts of perceiving, acts of recognizing, acts of proving and acts of insight are examples of the former. An act of insight or understanding may be expressed—'Now I know it!'—or described—'Suddenly I knew.' These acts are allowed to have nonhomogenous temporal parts. A state of certainty is preferably not called a state of knowing. Knowledge may be accompanied by a state of (subjective) certainty, that is, a high degree of conviction, but the two concepts should be distinguished, for I may be in a state of certainty without knowing, and I may be knowing without being in a state of subjective certainty. Knowing is thus not a state in the sense in which a state of doubt is a state.

Knowing whom Prince William had married, what the sum of 7 and 5 is and that it has been raining yesterday are examples of capacities that can be actualized in acts of judging. We are said to know these things whether awake or asleep. The boy knows the sum of 7 and 5 as soon as he has understood that the sum of 7 and 5 is 12, and has stored this in his memory in such a way that he is able to judge again that 7 and 5 is 12. Understanding that 7 and 5 is 12 is an act of knowing, and the knowledge obtained through this act of understanding is a specific capacity. The boy is able to actualize such a capacity, and will do so when asked by the teacher, by making the assertion that 7 plus 5 is 12.

Following the Aristotelian order of explanation, the first way to relate the concepts knowledge and belief can be presented in two ways. Knowledge as capacity is explained in terms of belief as a capacity to judge, and belief is explained in terms of the act of judgement. Or, knowledge is explained in terms of the act of knowing, and the act of knowing can then be explained in terms of the act of judging.

Kevin Mulligan has given an argument to understand knowledge, including judgement, and belief to be of different categories, which can be used as an argument for the thesis that belief, on the one hand, and judgement and knowledge, on the other hand, belong to different categories. 'She doesn't believe that *S*' is ambiguous, meaning either the same as 'It is not the case that she believes that *S*', or the same as 'She believes that it is not the case that *S*'. The verb 'believe' is thus a Neg-Raiser (cf. Mulligan 2013). The sentence 'She doesn't know/judge that *S*' has only one meaning, for it precisely means 'It is not the case that she knows/judges that *S*'. Another difference between judgement and knowledge, on the one hand, and belief, on the other hand, consists in the fact that belief has a negative counterpart, disbelief (cf. Mulligan 2013). Judgement and knowledge do not have such negative counterparts. Or, if judgement has a negative counterpart in the form of a rejection, it seems rather that a former assertion or proposition that is put forward is rejected. Such a rejection, though, always implies an act of judgement or assertion (cf. Reinach 1911, p. 365). For, even a rejection has assertive force. This means that rejection is not the negative counterpart of judgement, and that rejection and judgement are not concepts on the same level. The notions of belief and disbelief, in contrast, are on the same level, at least, if belief and disbelief are understood as different degrees of conviction.[4]

(ii) It is therefore not unlikely that knowledge and belief are not to be explained in terms of each other; they seem to be exclusive categories. When someone asks me whether I believe that John is unfaithful to his wife, I might answer: 'I do not believe it; I *know* it.' The Platonic distinction between scientific knowledge (*episteme*) and opinion (*doxa*) is not only exclusive, knowledge and opinion also have different objects. On the Platonic account, 'belief' means mere opinion.

A contrast between the terms 'knowledge' and 'belief' may also be used as an expression of the contrast between the concepts knowledge and unquestioned faith. In Wittgenstein's (1951) *On Certainty,* one can find the idea that we have

[4] '[P]ositive and negative conviction are ranged alongside each other on an equal footing' Reinach (1911, p. 333). The original has: 'Positive und negative Überzeugung stehen,…, einander gleichgeordnet gegenüber Cf. Mulligan (1987).'

an unquestioned faith in certain propositions, the hinge propositions, and that this makes it possible that other propositions may be doubted, or be certain and known. Such an unquestioned faith is improperly expressed by a declarative sentence, for hinge propositions are neither true nor false (§ 205). Unquestioned faith rather shows itself in the way we act (§ 402). Such a faith is hinted at by Husserl when he speaks of 'das Weltglauben', our belief in the being of the world, which never will be doubted (cf. Husserl 1939/1985, § 7, p. 25).

We also find a contrast between knowledge and belief in Hume's writings. Knowledge, in *A Treatise of Human Nature* (1739/2009), is concerned only with relations of ideas, and is therefore certain, whereas belief's objects are matters of fact, which means that our belief may at most reach a certain degree of probability. Belief is thus not a general category of which knowledge is a species; knowledge is certain, whereas belief is probable. In this sense, knowledge and belief are exclusive categories, and belief has the meaning of opinion. This is not the only meaning of 'belief' in Hume's writings. There are degrees of belief (Hume 1748/1999, *Enquiry,* § 6, p. 131), and because Hume considers belief to be a certain feeling or sentiment (*Enquiry,* § 5, part 2, p. 124), namely of security, 'belief' also has the meaning of state of conviction. Furthermore, belief in external objects and an external universe is rather a 'natural instinct' (*Enquiry,* § 5, part 1, 123, and § 12, p. 200); 'belief' thus also means unquestioned faith. Furthermore, Hume aims to give an analysis of what is called 'judgement' in the tradition. In Hume's mental geography, belief or judgement is primarily an act of the mind (*Treatise* 1.3.7, 67, note 20), that is, an act of judgement. Hume attacks the traditional account of judgement that he attributes to Locke. Hume has in this respect more in common, though, with Locke than he wants us to believe. Like Locke, Hume considers knowledge and belief to be exclusive categories, and he likewise exploits all the ambiguities of the term 'belief' and 'judgement'. For Locke, though, the faculty of judgement is, like the faculty of knowledge, a rational one, and it is therefore possible for him to give an explanation of judgement that is analogous to that of knowledge (on Locke's use of the terms 'belief' and 'judgement', see van der Schaar 2008). Hume thus separates the concepts knowledge and belief in a more radical way than Locke has done.

One may wonder whether knowledge and belief share a common genus. Do both concepts involve the concept of a capacity to judge? Is this capacity to judge a *genus,* of which knowledge and belief are two different species? This is not the way Locke or Hume conceive the two notions: Knowledge as act is for Locke a primitive notion, not to be explained in terms of any other notion. And knowledge as capacity is explained in terms of the act of knowing. For Locke, the act of judgement is a primitive notion, too, elucidated in analogy with the act of knowing, but not explained in terms of 'knowing'.

Knowledge and belief also seem to be of another category because we form different types of question with each of these notions, as reaction to an assertion. If John asserts that it rains, a proper reaction is to ask: 'Why do you believe/judge that it rains?' If the interlocutor himself knows that it rains, he is also entitled to ask: 'How do you know that it rains?' The 'How' question only makes sense if the agent who asks the question takes himself to know that it rains, and the answer

should describe the *one* way through which the agent who made the assertion came to know the fact. The 'How' question points to the fact that we do take knowledge to be a product of an act of coming to know, and that we consider this act of coming to know to be of importance for the judger's knowledge and reliability. The 'Why do you believe that?' question is 'a request for one's reasons for believing, or evidence for its being the case' (Hacker 2004, p. 217). In answer to the 'Why' question, one may give a *series* of reasons. There is a third type of question that we sometimes ask when an assertion is made: 'What is the cause of your conviction?', or 'To what/whom do you owe your conviction?'. For example, when the assertion is made 'Without Churchill England would have been defeated by Hitler', one may ask 'What is the cause of your conviction?'[5] The answer may be: 'It is due to my primary school teacher, Mr. Higgins, for whom Churchill was a hero.' It is precisely this type of question that Hume addresses in his science of the mind: 'What are the causes of our beliefs/convictions about the near future?' Convictions stand in relations of cause and effect, and it is this relation that is addressed in this type of question.

(iii) A third way to relate the concepts of knowledge and belief is to explain belief in terms of knowledge. This means that knowledge is a concept prior in the order of explanation to belief. Not all meanings of 'belief' distinguished above seem to be relevant here: Mere belief or opinion is pre-eminently secondary in the order of explanation to knowledge. With the benefit of hindsight one might say: 'I used to think I knew this, but I now see that it was mere belief.' A judgement or assertion purports, or is expected, to be knowledge, or at least true, but may turn out not to be what it purports to be, at a later time. If someone asserted this morning that it would rain today, and it turns out that it did rain, we assume that he or she did know it already this morning. If it turns out that it did not rain, he may say: 'Sorry, I believed so, but I was wrong.' If it is not a fact that it has been raining today, one cannot know that it rains today. The most we can say is that the asserter believed that it would rain today. The moment it becomes clear to the agent that it has not been raining today, he is expected to withdraw his assertion, and the corresponding belief will disappear. One may speak about the mere judgement or mere belief one had only with hindsight knowledge.

There is here an agreement with speech acts that are misfires. Just as an utterance like 'I name this ship the Lady Di' can be a misfire, because 'the procedure that we purport to invoke is disallowed or is *botched*' (Austin 1962, p. 16; italics mine), in this case, because the speaker is not entitled to name the ship, so a mental act like a judgement can be a misfire because it does not do what it purports to do. Like the botched speech act, mere belief or mere judgement is void or without effect. The word 'mere' in 'mere belief' is not enriching the meaning of the term 'belief' in the way attributive terms do. A common belief is a belief that is commonly held. The

[5] The Dutch language has two question-words related to the two meanings of 'cause': (1) cause as reason, which may be asked for by a 'waarom' question, the 'Why' question mentioned above ('Waarom geloof je dat?'); and (2) a metaphysical cause, which has the notion of effect as a counterpart. A cause in this sense may be asked for by a 'waardoor' question ('Waardoor heb je die overtuiging?', 'What is the cause of your conviction?').

term ‘common’ is thus enriching the meaning of the predicate: The extension of the phrase ‘common belief’ is therefore smaller than that of ‘belief’. Commonly held beliefs form a subclass of beliefs. In its modern meaning, ‘mere’ is synonymous with ‘nothing but’. A ‘mere belief’ is a belief that is nothing but belief, that is, a belief that is not what it originally purported to be, namely knowledge. Such a term as ‘mere’ can be informative, though: Asserting that something is ‘mere belief’ is asserting that it is a belief, and that it is not what it purports to be, which, if true, is more informative than asserting that it is a belief.

At first sight, one might think that mere belief is a species of belief, but the problem is that mere belief does not have a specific difference that distinguishes it from other beliefs, for these other beliefs may also turn out not to be what they purport to be. The relation between the concepts belief and mere belief seems to be unique, not unlike the relation between the notions water and pure water. The terms ‘mere’ and ‘pure’, and other nonattributive terms, such as ‘botched’, are important in order to understand our epistemic language. Section 11.3 will be devoted to these etiolations of language, as Austin calls them, and in the last section I will show how these terms can be of use in the explanation of mere belief as botched knowing. Standard forms of conceptual analysis cannot be of use in the elucidation of the relation between belief and mere belief; a different form of conceptual analysis needs to be developed in the sections below.

(iv) The explanation of concepts such as knowledge, belief, judgement, assertion, justification or ground and truth will always be circular, one might argue, because none of the attempts (i), (ii) and (iii) give an ultimate account of these notions. One may give necessary and sufficient conditions for knowledge, but the terms used in these conditions are not better understood than the concept of knowledge itself. To understand these notions is precisely to understand how these concepts are related to each other. None of these notions is clearer than any of the others, which means that none can be used to give an explanation of these notions.

The danger of this position is that it may function as a licence to leave the concepts as they stand. When one recognizes that the term ‘belief’ has several meanings, it is possible to understand that when one explains knowledge in terms of belief, one makes use of a certain meaning of the term ‘belief’, namely that of a certain degree of conviction, and, perhaps, of a capacity to judge, and that, when one explains belief as botched knowing, one makes use of the concept of mere belief or opinion. Perhaps, those who think that only a circular explanation can be given of these notions, do not realize that the meaning of their terms changes in the different accounts given.

11.3 Modification and Other Etiolations of Language

Modern philosophy has been interested in three of the four meanings of the term ‘belief’: judgement and its linguistic counterpart assertion; conviction and degrees of belief; and unquestioned faith, whether in the form of religious belief, animal

instinct or faith in hinge propositions. The concept of mere belief or opinion has not received much attention (Price 1969 is an exception): The idea that mere belief is botched knowing is hinted at, but not worked out. The idea is often used, though, in everyday language when people speak of 'subjective opinions'; in Dutch one often hears, 'dat is maar een mening' ('that is nothing but an opinion', in English), implying that it is not knowledge, and that all opinions are of equal epistemic value. Early analytic philosophy conceives complex concepts to be wholes consisting of atomic parts, which can be obtained by analysing the complex concept.[6] Or, it conceives of concepts as obtained through analysis of judgemental contents, thus giving different concepts depending on the way the judgemental content is analysed (Frege 1879/1971, § 9). In both cases, the concept of mere belief cannot be further analysed, and cannot be elucidated by means of the concept belief. The relation between the concepts belief and mere belief should therefore be elucidated in a nonstandard way.

Mere belief is not a special case of belief in the way a true or a false belief is, but it is certainly a belief. Belief purports to be knowledge, or, at least, to be true. If one finds out that the content of the belief that someone holds is ungrounded or false, one may call the belief 'mere belief', that is, a belief that is not what it purports to be. We first have to understand that a belief is expected to be knowledge in order to determine what 'mere belief' means. The complex term 'mere belief' is thus parasitic upon the normal use of the term 'belief'. We can use the term 'mere' in combination with other nouns: Her mere presence makes him angry. One expects that the things she does make him angry, but these expectations are unfulfilled.

'Mere' does not have an independent meaning; it is a syncategorematic term, a term that can be given a meaning only together with another term, a noun, to which it belongs. The term 'mere' functions as an operator upon the term that follows, but not precisely in the way negation does. Unlike propositional negation, 'mere' cannot sensibly be iterated. It is true that the term 'mere' includes a negation, for it means 'and nothing more', but what is negated is not the attribute that is denoted by the general term. For a mere judgement is a judgement. What is negated is determined by what we expect if something is called a 'judgement'. The term 'mere' restricts the meaning of 'belief', 'presence' or whatever noun that follows, in the sense that it denies a purported or expected aspect of what is denoted by the noun. For this reason, I call 'mere' a *restrictive* term. Restrictive terms are not attributive: We do not use them to attribute a property to an object. In standard cases, adjectival or adverbial terms are attributive in the sense that the speaker attributes a property to an object. The term 'rainy' is attributively used in 'It is a rainy night.' This means that one can use the sentence as a premise and draw the conclusion 'It is rainy and it is night.' The term 'mere' is nonattributive for one can say 'That is mere belief', but one cannot say 'That is a belief that is mere'.

The term 'pure' has characteristics similar to the term 'mere'. If we say that a statue is made of pure gold, we say that it is made of gold and of nothing else. A pure-bred Arabian is an Arabian horse that is not mixed with other breeds. The

[6] 'A thing becomes intelligible first when it is analysed into its constituent concepts.' Moore (1899, p. 182).

difference is one of evaluation: The term ‘mere’ can generally be substituted by the phrase ‘nothing more than’ whereas the term ‘pure’ can be substituted by ‘nothing less than’, ‘not mixed with anything else, especially things of lesser value’, as in ‘pure wine’. Sometimes we can use both ‘pure’ and ‘mere’: ‘pure Platonic love’ or ‘mere Platonic love’, depending on what is given a higher value. ‘Pure’ is a restrictive term, too.

In the paper ‘Wooden Horses and False Friends: On the Logic of Adjectives’ (van der Schaar 2014), I have given a classification of nonattributive terms inspired by Twardowski’s three-page paper from 1923. A term *A* is attributive precisely if the following inference is valid; otherwise, it is nonattributive:

I. *a* is (an) *A N*.

$\vdash$ *a* is *A* and *a* is (an) *N*.

If it is a red jacket, it follows that it is red and a jacket. ‘Red’ is therefore an attributive term. Relative terms such as ‘good’ and ‘big’ are on this account nonattributive, for a good thief might not be good (as a man). Apart from the relative terms, there are two kinds of nonattributive terms, the modifying terms such as ‘alleged’, ‘fake’, ‘botched’ and ‘non’, and the non-modifying ones, such as ‘mere’, ‘pure’, ‘proper’ and ‘real’. Modifying terms delete a part of the meaning of the term they belong to in such a way that a fake gun is not a gun, and botched knowing is not knowing at all, although it has the outer form of a gun, in the first example, of knowing, in the second.

A is a modifying term, precisely when inference II is valid:[7]

II. *a* is (an) *A N*.

$\vdash$ *a* is not (an) *N*.

In the case of ‘mere’ and ‘real’, inference schemes I and II are invalid. II is invalid for if it is mere belief it is belief. I is not valid, because the phrase ‘It is mere’ does not make sense.

Among the nonattributive, non-modifying terms one may distinguish between the terms ‘mere’ and ‘pure’, on the one hand, and terms such as ‘real’, ‘true’, ’proper’ and ‘authentic’, on the other hand. From the premise that he is a true friend it follows that he is a friend, but it does not follow that he is true, for the latter term does not have an independent meaning when used in these kinds of contexts. These terms are distinguished from ‘mere’ and ‘pure’ insofar as they have a modifying counterpart: ‘real’ and ‘fake’; ‘true’ and ‘false’; ‘proper’ and ‘improper’; ‘authentic’ and ‘inauthentic’. When it is suggested that someone is a false friend, the answer

[7] In the paper I argue against Barbara Partee’s (2010) thesis that fake guns are a special case of guns, which means that on her account modifying or privative terms behave rather like relative terms.

might be 'No, he is a real friend'. To understand the meaning of 'real friend', one needs to understand the meaning of 'not being a real friend', that is, of 'being a false friend'. It is the negative use that wears the trousers, as Austin says (Austin 1962, p. 70). We first have to understand what it is for something to be not real gold, that is, to be a material that looks like gold but does not have the chemical properties that would make it into a piece of gold. Only against this background does the phrase 'real gold' get a meaning. Saying that this is not a real Rembrandt, is not denying that it is a painting, nor is one generally claiming that it is a forgery. The question is rather whether it is painted by Rembrandt, or by one of his pupils. Only against the background of this question is the assertion that it is a real Rembrandt given sense. The question might also have been whether it is a forgery or not, and then asserting that it is a real Rembrandt has a different meaning. According to Twardowski, terms such as 'real', 'true' and 'actual' may restore the change in meaning that was caused by such terms as 'fake', 'false' and 'former' (Twardowski 1923, p. 142), and he calls them *restorative* terms.[8] Earlier than Twardowski, Brentano's older student Anton Marty has written not only about modifying terms, but also about these restorative terms. According to Marty (1884, p. 52, 53), terms such as 'real' and 'true' are able to restore the original meaning of the head noun, when the meaning has been modified by a term such as 'painted' in 'painted lion,' and 'past' and 'future'. Marty points to the fact that we sometimes can delete the modifying term, but still use the term in its modified meaning, for example, when we refer to a painting of a landscape by calling it 'a landscape'. Philosophical interest in these modifying terms in the nineteenth century originates with Brentano and Bolzano.[9]

In the first part of the *Wissenschaftslehre,* there are at least four places where Bolzano makes philosophical use of modifying terms (Bolzano 1837, §§ 19, 23, 29 and 59), although Bolzano does not call them that way. In Sect. 19, an interesting quote from the fifteenth-century Florentine Savonarola is given: '[Just as] dead man has the form and the likeness of a man, but is not a man' ([*Sicut*] *homo mortuus habet figuram et similitudinem hominis, non tamen est homo*). The explanation Savonarola gives is of interest, for it is true that if *A* is a modifying term and *N* the noun phrase, we can draw the conclusion that the object is not an *N,* and also that the object has the appearance or form of an *N* (Schaar 2014). Bolzano gives an explanation of phrases in which modifying adjectives occur. In standard cases such as 'golden candle', the head noun of the phrase indicates the head presentation of the complex presentation that is meant by the phrase. This means that a golden candle is a kind of candle. In cases where the adjective is a modifying term, it is not the head noun that indicates the head presentation, for example, in 'painted fish' the head noun 'fish' does not indicate the head presentation, for a painted fish is not a kind of fish. It is rather a kind of painting, which means that 'painted' indicates the head presentation (Bolzano 1837, I, § 59). This distinction can be used to clarify misleading terminology in philosophy. According to Bolzano, a possible thought, *cogitatio*

[8] I prefer this terminology to the less apt 'redundant terms': these terms are not redundant at all when used in the right context.

[9] Markus Stepanians and Arianna Betti have already pointed to the Bolzano side of the topic, cf. Stepanians (1998, p. 28) and Betti (2006, p. 59).

possibilis, is not a kind of thought, it is rather a kind of possibility (idem, § 23). Just as an apparent truth is not a kind of truth, so a relative truth is not a special kind of truth, either (idem, § 29). There are, according to Bolzano, also phrases where neither of the terms can be understood as indicating the head presentation. In such cases, we have an improper nomination (*uneigentliche Benennung,* idem § 182). Bolzano gives the example of 'a formal truth'. A formal truth is, for Bolzano, neither a kind of truth nor a kind of form. The point is that there are not two kinds of truths: the formal ones and the real ones. What philosophers call a formal truth is a proposition that does not contain a contradiction, but this need not be a truth at all. These philosophers take the term 'truth' thus too broad in extension (idem, § 29, p. 138, 139).

Brentano mentions the modifying terms in his logic lectures from the early seventies (Brentano 1870/2011, EL80-13.063[4]). When dealing with different kinds of ambiguities, he says that an adjective in the composition of an adjective and a substantive is normally determining, but may also be modifying, as in 'falsches Geld' and 'gedachter Taler'. This is not a true form of ambiguity, Brentano says, but rather an ambiguous form (*äquivoker Form*). The passage from the manuscript consists merely of notes and does not contain full sentences, so one has to give an interpretation. I assume that Brentano is pointing to a syntactic ambiguity: The combination of adjective and substantive is given a different meaning depending on the question whether the adjective is a modifying or a non-modifying term. So, there is a syntactic ambiguity in the phrase 'painted landscape', because the adjective may be understood as giving a determining quality to what is denoted by the noun, or the noun may be understood as qualifying what is referred to by the adjective: The thing that is painted is a landscape. Understanding modifying adjectives in the context of a syntactic ambiguity shows a certain similarity with Bolzano's analysis, although Brentano does not use Bolzano's concept of head presentation. Brentano's first publication on the topic can be found in his *Psychologie vom empirischen Standpunkt* (1874/1925, pp. 60–63). In a long note, Brentano quotes from the 'brouillon' of a letter he has sent to J.S. Mill, discussing Mill's thesis that the copula does not necessarily include existence. The correspondence is shortly before the publication of the *Psychology,* as Mill's answer is from February 1873. According to Mill, the sentence *A centaur is a fiction of the poets* does not include an affirmation of existence, for it cannot be implied that a centaur, being a fiction of the poets, exists (Mill 1843, Bk. I, Ch. IV, § 1). Brentano answers that the sentence does contain an affirmation of existence, for the sentence has the same meaning as 'There is a poetic fiction, in which the upper part of a human body and the trunk of a horse are unified in thought into a living being.' So, the main phrase is: 'there is a poetic fiction'. Again, not unlike Bolzano's analysis, who equally would say that the sentence 'There is a fictitious centaur' contains an assertion about a certain kind of fiction. There is no affinity, though, in terminology between Brentano and Bolzano. The truth of the sentence does not demand that there exists a centaur, but only that a fictitious centaur exists. The logical role of a term such as 'fictitious centaur' can be explained, when we understand that not all adjectives add something to the meaning of the term they belong to. Equally, the term 'dead' in the sentence 'a man

(*Mensch*) is dead' is a modifying term, for a dead man is not a man at all (Brentano 1874/1925, p. 62, note). Brentano understands that terms in predicative position may also function as modifying terms. The sentence shows its logical structure when it is rewritten as 'there exists a dead man'. Brentano uses here Savonarola's example, but there is no reason to presuppose that he got the example from Bolzano, for we find it in Aristotle's *De Interpretatione* (21a18 ff.), and it was well known in Scholastic literature (see below). Linguistic distinctions should not be confused with distinctions in thought, that is, with logical distinctions, Brentano ends the long note. In 1914, Brentano comes back to the modifying terms. Adding the word 'past', 'painted', 'thought' or 'apparent' to a noun, such as 'king', does not enrich, but rather modifies the meaning, deleting the original meaning and substituting another meaning instead. A past king as such is no more a king than a beggar is (Brentano 1914/1968, p. 46). The 1914 explanation is clearly different from the one given by Bolzano, as it is now given purely in semantic terms. Although the semantic explanation is published only in 1914, it is reasonable to assume that Brentano had used it in his lectures, and that Marty and Twardowski knew this explanation because they have attended these lectures, or have discussed the topic with Brentano in private.

Twardowski makes use of the distinction between attributive and modifying terms to explain the ambiguity of the term 'presented object'.[10] Twardowski's analysis is closer to Brentano's than to Bolzano's analysis. Twardowski refers in this context only to Brentano, and the explanation is given in semantic terms. According to Twardowski, an adjunctive clause is modifying, when it changes the original meaning of the name, to which the clause belongs, and substitutes a new meaning (Twardowski 1894, § 4, p. 13). Marty introduces the idea of modifying and restorative terms in the context of Mill's example and the Brentanian thesis that all judgements are existential in form. Like Brentano, he gives a semantic explanation of modifying terms: They 'modify' the meaning of the noun phrase (Marty 1884, p. 50). The source for Marty and Twardowski thus seems to be Brentano rather than Bolzano.

It is most likely that Bolzano and Brentano were familiar with the function of terms such as 'dead' and 'in thought' because Aristotle introduces them at several places, and because the function of these terms was familiar in the work of, for example, Ockham and Cajetan. Aristotle's *Categories* opens with the ambiguity or homonymy of certain names. The example he gives relates to the fact that in Greek the same name is used for a painting and for a man. Because a different definition is given for the two objects, Aristotle says, this is an example of homonymy. The example is not related to modifying terms, for there is a real ambiguity in the Greek

[10] Twardowski (1894, § 4, p. 12 ff.) uses the distinction between the modifying and the attributive sense of the term 'presented' to explain the distinction between the content and the object of an act. There is a distinction between presented object as object, where 'presented' is used in its attributive sense, because we say about the object that we have a presentation of it, and presented object as content, where 'presented' is taken in its modifying sense, because the term modifies the meaning of the term 'object'. If 'painted' is used as modifying term in 'painted landscape', the landscape is a painted one, that is, not a true landscape ('sie ist keine wahrhafte Landschaft,' Twardowski 1894, p. 13). 'Painted' can also be used attributively: We can talk of a landscape near Amsterdam that was painted by Rembrandt.

language. Aristotle comes back, though, to the topic of homonymy in a more interesting context in *De Anima* (412b17 ff.): 'The eye is matter for sight, and if this fails it is no longer an eye, except homonymously, just like an eye in stone or a painted eye.' (Hamlyn 1993, p. 10). If the word 'eye' refers to an eye that can no longer be used to see, the meaning of the term is different from the standard term, as in the case when it is followed by modifying phrases such as 'in stone' and 'painted'. The most important passage can be found, though, in *De Interpretatione* (21a18 ff.), the end of Chap. 11. Paraphrasing the first part of the passage: We may say if something is a white man, that he is white, and that he is a man. This does not hold for all terms, though. When the adjective contains something that is opposite to what is contained in the noun, a contradiction follows, and the conclusion is always false. For example, when we call a dead man a man, this is false, because the adjective *dead* contains the meaning of not living, whereas the noun *man* contains the meaning of living being. In terms of the validity scheme and our concept of modifying terms: 'Dead' is a modifying term, for if something is a dead man, it follows that it is not a man. But there are also cases, Aristotle adds, in which the conclusion may be true or false, that is, in which case it is invalid that if it is an *A N,* it is an *N,* as in the example of Homer is a poet, in which a poet qualifies the verb *is,* which means that the conclusion that Homer exists cannot be drawn. Aristotle ends the passage with an interesting example. If something is thought, or exists in thought, it does not follow that it is, or that it exists.[11] In the terminology introduced above, '(in) thought' behaves as a nonattributive term. Although we would now say that existence is not an ordinary predicate, it *is* a second order predicate, and may thus be qualified, on both the Aristotelian and the modern account. From the fact that something exists in thought, it cannot be inferred that it exists.

11.4 A Linguistic Phenomenology for Mere Belief

Sometimes it is said that belief or judgement aims at knowledge, or, at least, at truth.[12] This is a metaphorical way of speaking, for beliefs do not literally aim at something. What is meant can be formulated in two ways. From a first-person perspective, one can say that there is no distinction between belief and knowledge, because one takes one's beliefs to be knowledge.[13] This is not to say that the degree of confidence is the same in the case of belief as in the case of knowledge, although this may also be so. 'Belief' is here to be understood not as conviction, but as judge-

[11] Weidemann translates the example as follows: 'Von etwas Nichtseiendem kann aber nicht deshalb, weil es ein vermeintliches (Seiendes) ist, wahrheitsgemäss ausgesagt werden, es sei ein Seiendes. Denn die (blosse) Meinung, (es sei), hat man von ihm ja nicht etwa deshalb, weil es (seiend) wäre, sondern gerade deshalb, weil es nicht(seiend) ist.' Weidemann (1994, p. 24, 25).

[12] Pascal Engel, Truth, Acumen, 2002, p. 128.

[13] Cf. van der Schaar (2011a) and Adler (2002, p. 275): 'From the first-person point of view, one treats one's belief as factive, which is the central property of knowledge.'

ment, either in the act sense or in the sense of capacity, depending on whether we speak of acts of knowing or knowledge as (specific) capacity. From a third-person perspective, if a man utters a declarative, one takes him to make an assertion unless there are signs to the contrary. When one makes an assertion, one is expected to know what one asserts. For, an interlocutor is entitled to ask 'How do you know that?' When it is shown that what is asserted is false or ungrounded, that is, when it is shown that someone does not know what he has asserted, he is expected to withdraw his assertion. We take an assertion to be the manifestation of knowledge. The assertion, insofar as it is an outer act of judgement, is the actualization of knowledge as capacity.[14] From both the first- and the third-person point of view, there is thus a conceptual relation between assertion or judgement and knowledge.

It is a fact, though, that not all our assertions are manifestations of knowledge. Some of our judgements turn out to be mere judgements, and may therefore be called opinions. The terms 'mere judgement' and 'judgement' differ in meaning. The term 'mere' in front of the term 'judgement' works as an operator upon the meaning of the latter term, but not in the way a modifying term does, for a mere judgement is a judgement. If you call it a 'mere judgement', you are not denying that it is a judgement, but you are implicitly denying something. The term indicates that the denoted judgement is not what it purports to be, not what we expect it to be, namely knowledge, or true judgement, if you prefer. 'Mere judgement' means in this context (1) a judgemental act that purports, is expected, to be knowledge, but (2) is, from a third-person perspective, not knowledge. This concept of mere judgement thus has the concept of knowledge as part of its explanation. Because of the partial negation, the operator 'mere' restricts the meaning of the term 'judgement'. The term 'mere judgement' has the same meaning as 'opinion', in the sense in which it is opposed to knowledge, whose meaning can be stressed by speaking of 'mere opinion'. This means that 'act of mere judgement' has the same meaning as 'mere opining', the latter being a term that is not so common today, but in Dutch and German one still uses the terms 'menen' (Dutch) and 'meinen' (German) in the sense of *to opine*. In special cases, one may qualify one's own speech act as mere opinion. When a scientist is asked about his ideas on a topic, and he wants to stress the fact that he does not have knowledge in that field, he might put forward his ideas, and add: 'This is mere opinion, nothing more than my opinion'. Just as he may qualify his speech act by directly adding the parenthesis 'I think'.

According to Husserl in the fifth *Logical Investigation,* the term 'mere' (*das 'bloss'*) is a sign that there is a lack, a deficiency (*ein Mangel*), and this shows something about the order of explanation of the relevant concepts. Husserl uses this idea to elucidate the relation between the concept of perception and that of imagination, which can in this case be understood as a special case of the distinction between knowledge and mere belief.[15] The concept of perception is not obtained by adding something to the concept of mere imagination. The term 'mere' is a sign that

[14] I have given a full account of assertion in van der Schaar (2011b).

[15] 'Bloss (die Blösse) weist hier, wie überhaupt, auf einen Mangel hin; aber nicht immer ist ein Mangel durch eine Ergänzung zu beheben. So setzen wir ja der Wahrnehmung die "blosse" Ein-

the concept of perception is prior in the order of explanation, and that the concept of (mere) imagination is obtained by subtracting a part of the concept of perception. Husserl's example takes us to a waxwork show. When we enter the show, we see a charming, unknown lady on the staircase inviting us to come with her. One moment later, we realize that it is an optical illusion that a trick was played upon us. Now, we see a wax figure that is presenting a lady. According to Husserl, the perception when we enter the show does not consist in an act in which something that is common to the perception and the imagination is presented, together with an act of perception, which contains the 'belief' moment. The act when we enter the show is nothing but the perception of the lady on the stairs. Later, we perceive a puppet that has the appearance of a lady (Husserl 1901/2009, V, § 27, p. 458, 459), and we may call the former 'perception' a mere imagination. The idea of a mere imagination of the lady on the stairs is to be explained in terms of a perception that is lacking something.

In *Experience and Judgement,* Husserl says that the phenomenological genesis of judgement shows us something about the order of explanation of the concepts mere judgement and knowing judgement. We first experience assertions as products of purported knowing (*prätendierte Erkenntnisse*), and we do not distinguish between mere judgement (*bloss prätendierte, blosse Urteile*) and knowing judgement (*wirkliche Erkenntnis*) (Husserl 1939/1985, § 5, p. 15). Only in a later phase, we may come to realize that the judgement is a mere judgement. This shows, according to Husserl, that mere judgement is 'an intentional modification' of knowing judgement.[16] The term 'mere' (*bloss*) is not a modifying term, so how should one read Husserl here? In the first place, Husserl does not understand the operation of modification in semantic terms: there is no term that does the operation of modifying. For Husserl, the modification is intentional in the sense that there is a change in the intentional content of the act in relation to the original intentional content, where the intentional content is a complex consisting of both the matter and the quality of the act. More specifically, there is a change in the quality of the act in relation to the original quality.

This type of modification is distinguished from semantic modification insofar as there is no term that operates as a modifier. The concept of mere judgement is obtained by deleting a part of the more primitive notion of purported knowing. If the change is a modification, something essential is missing from a mere judgement as compared to the knowing judgement, notwithstanding a similarity in form. The mere judgement is botched knowing, to use the terminology introduced earlier, and 'botched' *is* a modifying term, having 'proper' as its restorative counterpart. A proper judgement need not be knowledge, but it needs to be grounded in order to fulfil its function as purported knowing. As in the case of semantic modification, the concept of judgement as purported knowing is prior in the order of explanation

bildung gegenüber. Das Unterscheidende liegt in einem Vorzug auf Seiten der Wahrnehmung, aber nicht in einem Plus.' Husserl (1901, V, § 28, p. 463).

[16] '[D]ass blosses Urteilen eine intentionale Modifikation von erkennendem Urteilen ist.' (Husserl 1939, § 5, p. 15). Stepanians (1998, Chap. 10) gives an extensive account of Husserl's idea of modification in the fourth and fifth Logical Investigation.

to the result of the modification, the concept of mere judgement or botched knowing. And the restorative notion of proper judgement can only be obtained on the basis of the latter concept. One has to understand first what *purported knowing* is in order to understand what *mere judgement* or *botched knowing* is. The relation of modification between the two intentional contents is not to be understood in any psychological or biological sense, Husserl says in the *Logical Investigations* (Husserl 1901/2009, p. 488). The relation of modification expresses rather a relation between two essences. It shows something essential about mere judgement that it is understood as a modification of judgement as purported knowing. We may call this type of modification *conceptual*.

Besides semantic and conceptual modification, there seems to be a third type of modification, ontological modification, or what is generally called privation. A blind man is deprived of his ability to see, an ability that belongs to his humanity (cf. Aristotle's *Metaphysics,* book Δ; chapter V; section xxii; cf. *Met.* X. 1055a, 1980). Probably, Husserl was pointing to this type of modification when he claimed he did not understand modification in any biological or empirical sense. Blindness as privation is not mere absence. A stone does not have the ability to see, but it is not deprived of this ability; we do not speak of blind stones. The Augustinian tradition has it that evil is privation of the good.[17] And, because cognitive error is a special case of moral error, or sin, cognitive error is a privation, too. It is in this Augustinian sense that Descartes explains error as privation in the fourth *Meditation* (*privatio, sive carentia,* AT 7: 55).

It may be doubted, though, that we need a separate concept of ontological modification, besides conceptual modification. For Spinoza, error and sin are concepts that make sense only insofar as *we* compare things with one another, and this holds for everything we call a 'privation'. From God's point of view, there is no privation.[18] The false judgement may be called a 'privation' insofar as we compare the judgement with a knowing judgement; we can attribute the privation only insofar as we relate the false judgement to the idea of a knowing judgement. A mere belief is called 'botched knowing' only insofar as we expect our beliefs to be knowledge, and the concept of mere belief can thus be understood as the result of a conceptual modification, as Husserl has shown. This is not to say, though, that we can decide either to expect or not to expect our beliefs to be knowledge: Conceptual modification is not a psychological notion. It is part of the concept of judgement that judgement purports to be knowledge.

[17] As Augustine (1988) writes in the Confessiones, III. vii. 12: 'I did not know [at that time] that evil was only the privation of good'.

[18] '[W]hen we consider God's decree and God's nature, we can no more assert of that man that he is deprived of sight than we can assert it of a stone … privation is simply to deny of a thing something that we judge pertains to its nature,' Ep. 21 to Willem van Blyenbergh (Spinoza 1992). In the Cogitata Metaphysica (1982, Part II, Chap. 7), Spinoza says that evil and sin are nothing in things, but only in the human mind as it compares things with one another.

11.5 Conclusion

The idea that mere belief is a form of botched knowing can be used to elucidate the concept of mere belief as a conceptual or semantic modification of the concept of knowing, or purported knowing. The term 'botched' in 'botched knowing', being a modifying term, involves the idea that what is called botched knowing is denied to be knowing, although it does have the appearance of knowing, and may therefore be misleading. By allowing nonattributive terms in our logical geography, we are able to relate philosophical concepts such as knowing and believing in a more sophisticated way, on the presupposition that we have disentangled the ambiguities of our central philosophical terms.

References

Adler JE (2002) Belief's own ethics. MIT, Cambridge
Aristotle (1980) The metaphysics, vol I, II. Harvard University, Cambridge
Augustine (1988) Confessiones/Belijdenissen. Ambo, Baarn
Austin JL (1946) Other minds. In: Austin JL (1970) Philosophical papers, 2nd edn. Oxford University Press, Oxford, pp 76–116
Austin JL (1956) A plea for excuses. In: Austin JL (1970) Philosophical papers, 2nd edn. Oxford University Press, Oxford, pp 175–204
Austin JL (1962) How to do things with words. Oxford University Press, Oxford, 1984
Austin JL (1964) Sense and Sensibilia. Oxford University Press, Oxford
Betti A (2006) The strange case of Savonarola and the painted fish; on the bolzanization of Polish thought. In: Chrudzimski A, Lukasiwicz D (eds) Actions, products and things. Ontos, Frankfurt, pp 55–81
Bolzano B (1837/1929) Wissenschaftslehre. Felix Meiner, Leipzig
Brentano F (1870/2011) *Logik*. EL 80, In: Rollinger R. (ed) Salzburg. Available with: http://gandalf.uib.no/Brentano/texts/el/logik/norm/. Accessed 5 May 2013
Brentano F (1874/1925) Psychologie vom empirischen Standpunkt, II. Felix Meiner, Hamburg
Brentano F (1914/1968) Psychologie vom empirischen Standpunkt, III. Felix Meiner, Hamburg
Frege G (1879/1971) Begriffsschrift. In: Angelelli I (ed) Begriffsschrift und andere Aufsätze. Georg Olms, Hildesheim
Hacker PMS (2004) Of the ontology of belief. In: Siebel M, Textor M (eds) Semantik und Ontologie. Ontos, Frankfurt, pp 185–222
Hamlyn DW (1993) Aristotle De Anima books II and III. Clarendon, Oxford
Hume D (1739/2009) A treatise of human nature, vol I. Oxford University Press, Oxford
Hume D (1748/1999) An enquiry concerning human understanding. Oxford University Press, Oxford.
Husserl E (1901/2009) Logische Untersuchungen. Felix Meiner Verlag, Hamburg
Husserl E (1939/1985) Erfahrung und Urteil. Felix Meiner Verlag, Hamburg
Marty A (1884) Über subjektlose Sätze und das Verhältnis der Grammatik zu Logik und Psychologie, I. In: Marty A (1918) Gesammelte Schriften. Niemeyer, Halle, pp 3–101
Mill JS (1843) A System of Logic. In: Robson JM (ed) (1973/1974) Collected Works of John Stuart Mill, vol 7, 8. University of Toronto, Routledge
Moore GE (1899) The nature of judgment. Mind 8:176–193

Mulligan K (1987) Promisings and other social acts: their constituents and structure. In: Mulligan K (ed) Speech Act and Sachverhalt; Reinach and the Foundations of Realist Phenomenology. Martinus Nijhoff, Dordrecht, pp 29–90

Mulligan K (2013) Acceptance, acknowledgment, affirmation, agreement, assertion, belief, certainty, conviction, denial, judgment, refusal & rejection. In: Textor M (ed) Judgement and truth in early analytic philosophy and phenomenology. Palgrave Macmillan, Basingstoke, pp 97–136

Partee B (2010) Privative adjectives: subsective plus coercion. In: Bauerle R, Reyle U, Zimmerman TE (eds) Presuppositions and discourse: essays offered to Hans Kamp. Emerald, Bradford, pp 273–285

Price HH (1969) Belief. Allen & Unwin, London

Reinach A (1911) On the theory of the negative judgement. In: Smith B (ed) (1982) Parts and moments. Philosophia Verlag, München/Wien, pp 315–400

Spinoza B (1982) Cogitata Metaphysica/Metafysische Gedachten, Appendix to Spinoza's Descartes' principles of philosophy. In: Spinoza B Korte Geschriften. Wereldbibliotheek, Amsterdam

Spinoza B (1992) Correspondence/Briefwisseling. Wereldbibliotheek, Amsterdam. English edition: Spinoza B (1995) The Letters (trans: Samuel S). Hackett, Indianapolis

Stepanians MS (1998) Frege und Husserl über Urteilen und Denken. Ferdinand Schöningh, Paderborn

Twardowski K (1894) Zur Lehre vom Inhalt und Gegenstand der Vorstellungen Eine psychologische Untersuchung. Hölder, Wien. Second edition (1982) Philosophia Verlag, München, Wien

Twardowski K (1912) Actions and products. Some remarks from the borderline of psychology, grammar and logic. In: Twardowski (1999) 103–132

Twardowski K (1923) On the logic of adjectives in. Twardowski (1999), pp 141–143

Twardowski K (1999) Brandl J, Woleński J (eds) On actions, products and other topics in philosophy. Rodopi, Amsterdam

van der Schaar M (2007) The assertion-candidate and the meaning of mood. Synthese 159:61–82

van der Schaar M (2008) Locke and Arnauld on judgement and proposition. Hist Philos Logic 29:327–341

van der Schaar M (2009) Judgement, belief and acceptance. In: Primiero G, Rahman S (eds) Acts of knowledge: history, philosophy and logic; essays dedicated to Göran Sundholm. College Publications, London, pp 267–286

van der Schaar M (2011a) The cognitive act and the first-person perspective; an epistemology for constructive type theory. Synthese 180:391–417

van der Schaar M (2011b) Assertion and grounding; a theory of assertion for constructive type theory. Synthese 183:187–210

van der Schaar M (2014) Wooden horses and false friends; on the logic of adjectives. In: Mulligan K, Kijania-Placek K, Placek T (eds) The history and philosophy of polish logic; essays in honour of Jan Wolenski. Palgrave Macmillan, Basingstoke, 103–116

Weidemann H (1994) Aristoteles, Peri Hermeneias. Akademia Verlag, Berlin

Williamson T (2000) Knowledge and its limits. Clarendon, Oxford

Wittgenstein L (1951) Über Gewissheit. In: Wittgenstein L (1984) Werkausgabe 8. Suhrkamp, Frankfurt a. M.

Chapter 12
The Epistemological Disunity of Memory

Fabrice Teroni

Abstract A long-standing debate surrounds the question as to what justifies memory judgements. According to the Past Reason Theory (PastRT), these judgements are justified by the reasons we had to make identical judgements in the past, whereas the Present Reason Theory claims that these justifying reasons are to be found at the time we pass the memory judgements. In this chapter, I defend the original claim that, far from being exclusive, these two theories should be applied to different kinds of memory judgements. The PastRT offers the most appealing account of justified propositional memory judgements, while the Present Reason Theory provides the best approach to justified episodic memory judgements. One outcome of my discussion is thus that memory is not epistemologically unified and my argument in favour of this conclusion connects with the issues of internalism, reliabilism and the basing relation.

Keywords Memory · Episodic · Propositional · Justification · Impressions

The idea that our confidence in memory, which explains many if not most of the judgements we make, is systematically misplaced constitutes one of the most extreme and counterintuitive forms of scepticism. Intuitively, it seems obvious that memory judgements can be and typically are justified and this intuition has been happily endorsed by the vast majority of philosophers interested in the epistemology of memory. Yet, consensus ends here and the question as to what justifies memory judgements has generated a long-standing debate between two approaches. On the one hand, the 'Past Reason Theory' (PastRT) (e.g. Annis 1980, Malcolm 1963, Naylor 1985, Senor 1993) has it that memory judgements are justified by the reasons one had to make these judgements. What justify your memory judgements are those reasons you had to make them in the first place. On the other hand, the 'Present Reason Theory' (e.g. Audi 1995, Chisholm 1989, Ginet 1975, Pollock 1974) defends the claim that these reasons are to be found at the time memory judgements are made. What justify your memory judgements are reasons you have now that you remember. Is it possible to resolve this debate?

F. Teroni (✉)
University of Bern, Bern, Switzerland
e-mail: fabrice.teroni@philo.unibe.ch

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_12,

As opposed to previous treatments of this issue, my argument shall be that, far from being exclusive, these two approaches are needed to account for the justification of different kinds of memory judgements. For that reason, one outcome of my discussion is that memory judgements exhibit no epistemological unity. I shall proceed as follows. I first introduce the distinction between episodic and semantic-propositional memory and explain the epistemological import of some influent suggestions as to how we should draw this distinction. Next, in Sect. 2, I argue that the Present Reason Theory fails to account for the epistemology of propositional memory and suggests that this failure traces back to its relying on a distinctive form of internalism. This, I maintain, supports the PastRT as an account of propositional memory judgements. Yet, it would be wrong to think that this conclusion carries over to the epistemology of episodic memory. As I argue in Sect. 3, the Present Reason Theory offers here a much more sensible account. Finally, in the fourth section, I suggest a way of resolving the debate between the Past and Present Reason Theories.

12.1 Propositional and Episodic Memory

It is customary for psychologists (e.g. Tulving 1972, 1985, Perner 2000, Perner and Ruffman 1995) as well as for philosophers (e.g. Dokic 1997, Hoerl 2001, Martin 2001) to draw a contrast between propositional-semantic memory and episodic memory. The import of this distinction can be measured if we focus for a moment on two kinds of memory reports. 'John remembers that Dreyfus was innocent' and 'June remembers that Masaccio painted the *Cappella Brancacci*' are typical reports of propositional memory. In these reports, 'to remember' is followed by 'that' clauses and one of their salient features lies in the fact that they extend well beyond events the subject witnessed or objects with which he has been acquainted. Reports such as 'Mary remembers her first meeting with her boss' and 'Jim remembers an awful accident' are typical reports of episodic memory. In such reports, the verb is not followed by 'that' clauses. One salient feature of these reports consists in the fact that they are, as opposed to propositional memory reports, limited to events the subject has witnessed and objects with which he has been acquainted. This is why propositional memory reports are not easily turned into episodic memory reports. It would, for instance, be inappropriate to say of June, born in 1980, that she remembers Masaccio painting the *Cappella Brancacci*.

How should we draw the distinction between propositional and episodic memory? Here are three attempts to do so that have proved particularly influent. One approach has it that the presence of information about where and when something happened is the distinctive mark of episodic memory (Tulving 1972). Another approach suggests that what is specific of episodic memory is the presence of information about the self (Howe 2000). Finally, some have attempted to draw this distinction by appealing to the complex metarepresentational content referring to a past experience that is claimed to be distinctive of episodic memory (Dokic 1997, Perner and Ruffman 1995).

These claims about episodic memory will not be my main focus in what follows, but they help pin down two recurrent features of discussions of episodic memory that will prove important to disentangle some epistemological problems that lie ahead of us. First, the distinctiveness of episodic memory is claimed to reside in its subject matter—be it a past context, the self or a past experience. Episodic memory is (typically at least) constituted by a past-tensed judgement about the relevant subject matter. Second, these claims implicitly assume that propositional and episodic memories are epistemologically on a par. This may well be motivated by what I just said. For if episodic memory differs from propositional memory only insofar as the relevant judgements have a different subject matter, the difference may prove epistemologically inconsequential. This assumption can be questioned, however. And, as we shall see, investigating the epistemological issues surrounding memory provides some reasons to reject it. Let me start by propositional memory.

12.2 Propositional Memory

Propositional memory reports—'John remembers that Dreyfus was innocent'—have a clear epistemic import. We see John as, ceteris paribus, justified to judge that Dreyfus was innocent if he so remembers. Yet, in virtue of what are these memory judgements justified? I shall start by investigating how this question can be answered within the Present Reason Theory, which you will remember is the claim that the relevant reasons are to be found at the time memory judgements are made. If one subscribes to this theory, three ways of answering this question might be pursued: one may appeal to memory impressions, to some sort of inference or to the subject's beliefs about the source of his judgement. Let us see how these answers proceed, starting with the most intuitively appealing one.

12.2.1 Memory Impressions

We are all accustomed to the fact that judgements sometimes strike us as, to put it loosely, 'coming from the past'. 'It just seems to me that I remember that *p*', 'I have the impression that I remember that *p*' is how we typically express this fact. This appears to support a quite appealing idea: when one remembers that *p,* what justifies one in judging is one's seeming to remember that *p*. This claim is distinctive of what I shall call the *Memory Impression Theory* or, for short, MIT (Pollock 1974, Pollock and Cruz 1999, Chap. 2). This theory qualifies as an instance of the Present Reason Theory, since the memory impressions that justify memory judgements take place at the time these judgements are made.

Now, what exactly are these 'impressions' or 'seemings'? They should not be assimilated to judgements about our source for judging that *p* (Pollock 1974, p. 191). That is, the impression to remember that Dreyfus was innocent does not consist in John's additional judgement, e.g. that he learnt this in a serious book, but is a

distinctive phenomenological state. Let me grant this.[1] Next, note that if these impressions may be more or less intense, they are difficult to tell apart from one another. The seemings of seeming to remember that Dreyfus was innocent and seeming to remember that Napoleon crossed the Alps do not differ very much and this explains why advocates of MIT suggest that there is one sort of impression that simply gets attached to different contents.

The MIT offers an intuitively appealing account of justified propositional memory judgements, but it faces some significant difficulties. Let me start by observing that these memory impressions are often quite elusive. They do not accompany, for instance, many of our judgements about historical events. This is not in itself a problem for MIT, which can claim that these judgements are for that very reason unjustified. The problem is rather that the boundary between what is and what is not accompanied by these impressions hardly corresponds to the one we intuitively draw between justified and unjustified memory judgements. On the one hand, it would be preposterous to assess John's judgement that Dreyfus was innocent as unjustified simply because it is not accompanied by a memory impression. And, on the other hand, the presence of such an impression does not dispose us to think favourably of Max's judgement that aliens have landed in his backyard. Yet, Pollock argues that MIT has one advantage that might offset these serious misgivings:

> The recollection adds something to the belief. My having the recollection tags the source of the belief as being memory rather than present calculation or the result of reading the value off a table presently before me or simply pulling the number out of the air at random. (Pollock 1974, p. 189)

The argument is that memory impressions are needed in order for the subject to distinguish memory from other sources of belief or judgement. And it is true that, when trying to remember whether *p,* we sometimes eventually get a memory impression. Still, most propositional memory judgements arise spontaneously in answers to questions and are not prefaced by such tryings. And in these cases, no distinctive phenomenal impression is to be found: One spontaneously passes the judgement without being aware of a present source for making it. This is enough to be able to 'tag the source of the belief as being memory'. Distinguishing in propositional memory a memory impression that *p* and the judgement that *p* often does not correspond to any phenomenally salient fact (Locke 1971, p. 38; Naylor 1985).

We can now make use of these observations to build a dilemma. Either MIT persists in requiring memory impressions for justification, or it is modified so as to appeal to no more than the spontaneity of propositional memory judgements. The first horn is not attractive. It has, as we have seen, serious revisionist consequences regarding our intuitive grasp of the distinction between justified and unjustified memory judgements. The second horn is no more attractive. For the spontaneity of a judgement does not contribute in any way to its justification: Max's judgement that aliens have landed in his backyard remains unjustified despite its being spontaneously passed. The fact that many propositional memory judgements are made

[1] If you disagree, this means that MIT does not constitute an alternative to the Source Monitoring Theory I discuss below.

spontaneously does not help explain why they are justified. But MIT, appealing in the way it does to the circumstances surrounding the memory judgement, cannot but remain silent on what may account for the epistemological difference between two equally spontaneous judgements (see also Annis 1980, pp. 325–326).

Despite its initial attraction, then, an appeal to memory impressions fails to uncover necessary or sufficient conditions for the justification of propositional memory judgements.[2] Of course, MIT's dismissal does not signal the end of the Present Tense Theory, which comes in many different flavours. But if not memory impressions, then what sort of reason must we have at the time we remember in order for our propositional memory judgements to be justified?

12.2.2 Virtual Inferences

The second approach to propositional memory judgements suggests we answer this question by mentioning our capacity to back up these judgements by appealing to the evidence at our disposal. And it is definitively true that we are often able to back up these judgements in this way. To illustrate, let me come back to John. John, we may suppose, does not only remember that Dreyfus was innocent, but also that he just heard a serious historian telling so, and that he saw a BBC documentary arguing for this claim. This being the case, he is in a position to argue that the evidence at his disposal is best explained by the truth of his judgement. We can imagine him as arguing along the following lines: If Dreyfus had been guilty, then serious historians and BBC documentary makers would surely not be mistaken about it.

These observations may now be used to build an alternative epistemology of propositional memory judgements. Thus, Peacocke suggests that:

> A belief held without reasons is knowledge only if a sound, and in the circumstances knowledge-yielding, inference to the best explanation *could* be made from the evidence available to the believer to the truth of his belief. (1986, pp. 163–164)

Let me adapt this claim so as to make it apply to propositional memory judgements. The idea then becomes that these judgements are justified if a *justified* inference to the best explanation is at the believer's disposal. This inference being at the subject's disposal at the time he remembers, this explains why this *Virtual Inference Theory* (VIT) is, like the MIT, an instance of the Present Reason Theory.

In the light of the above observations about memory impressions, this approach has a lot to be said in its favour. Let me simply emphasize two important points. First, our attention is now drawn away from phenomenological issues and to the evidence at the subject's disposal, something which appears to carry more epistemological weight. Second, the spontaneous character of many propositional memory judgements is easily accounted for: VIT appeals to what is at the subject's disposal and does not require that he actually goes through a sophisticated inference.

[2] Senor (1993) and Owens (1999) offer further criticisms of MIT.

Yet, this last observation already contains the germ of a first difficulty faced by this approach. It is customary to distinguish between justifiable and justified judgements, i.e. between judgements we merely have reasons to make and judgements we make for these reasons. VIT is at pains to admit the existence of such a distinction for propositional memory judgements. Its appeal to the availability of an inference to the best explanation readily accounts for justifiable propositional memory judgements, but what about those that are in addition justified? It is certainly correct to point out that we should be ready to count among a subject's justified beliefs the obvious, though not drawn, consequences of what he justifiably believes. However, we cannot avail ourselves of this idea to extend VIT so as to account for justified memory judgements: it appeals to complex inferences, which are far from being obvious. This means that VIT must claim that propositional memory judgements are justified only if one *makes* the inference, i.e. that taking a reflective stance on doxastic states and their relations is required for justification. This gives rise to two difficulties.

First, the capacities that taking such a reflective stance presuppose have been quite consistently shown to be fully in place only around the age of five (e.g. Perner and Ruffman 1995). VIT thus unappealingly implies that no justified propositional memory judgement is possible before this age. The second difficulty has to do with the fact that we rarely indulge in the sort of inference around which this approach is built. Our reliance on memory is not shot through with inferences of this nature, which are made, or so it seems to me, almost exclusively when we nourish doubts about our beliefs. And it is not only questionable to try to build, for no obvious reason, the epistemology of propositional memory on psychological assumptions that are so rarely satisfied, one also cannot in this way account for all those judgements we intuitively assess as justified despite their being obviously not based on inferences.

Let me finally focus on another problem faced by the approach under discussion, a problem which centres on those quite common cases of propositional memory in which one has lost track of one's unique evidence for the relevant judgement. In these cases, there is not much by way of available inference. Suppose, for instance, that John once read that Dreyfus was innocent in a serious book and completely forgot this. Which inference can he then perform? To put it a bit oddly, the only evidence at his disposal is constituted by his belief. And even if we may try to apply VIT by suggesting that John's judging that his belief is best explained by its truth qualifies as an inference of the relevant type, this appears to be a constitutive feature of firmly held beliefs. This means that John is, according to VIT, on a par as regards justification with Mark, who I shall suppose acquired the same belief for fanciful reasons. But, surely, John differs from Mark with respect to justification. VIT is, we now realize, similar to MIT as regards the way it distorts our intuitive grasp of the domain. And it distorts it for the same reason, namely because it exclusively focuses

on what happens at the time John and Mark make their respective memory judgements, a time at which nothing tell them apart.[3]

All in all, then, the VIT does not prove more convincing than an appeal to memory impressions. At this stage, it seems to me that there is but one interesting option left for trying to account for propositional memory judgements within the boundaries fixed by the Present Reason Theory.

12.2.3 Source Monitoring

The key thought behind this third approach is constituted by the observation that, when we make propositional memory judgements, we are often able to back up our claims by reference to a source of information. John may for instance judge that he learnt that Dreyfus was innocent by reading a serious book: from his perspective, this is why he is now in a position to judge that Dreyfus was innocent. The distinctive claim of the *Source-Monitoring Theory* (SMT) is that propositional memory judgements are justified by further judgements of this nature. This is a third instance of the Present Reason Theory, because these source-monitoring judgements happen when memory judgements are passed. Yet, it differs from the MIT insofar as it does not rely on memory impressions, and from the VIT because no reference is made to complex inferences. It is nevertheless at least as problematic as these other forms of the Present Reason Theory.

The first problem is that SMT is, as VIT, cognitively quite demanding, since it makes justification depend on the deployment of source-monitoring capacities. For that reason, it does not allow us to draw a distinction between justified and unjustified propositional memory judgements for unsophisticated subjects. Insofar as we think that such a distinction can and should be made, SMT is a nonstarter.

The second problem also echoes one we have seen is faced by VIT. SMT relies on quite questionable psychological assumptions and as a result has some dire epistemological consequences. For note that we often do not preserve the source of our propositional memory judgements and that, when we are able to cite one, this is often rather due to reconstruction than to preservation. So, if SMT requires that the source be preserved, then it implies that many propositional memory judgements lack justification—such is the case for John's judgement, for instance, if he did not preserve its source and even though it manifests a belief based on the reading of a serious book. As a result, a large bulk of these judgements is left in the sceptic's hands. And SMT cannot appeal to reconstruction as such. Not only is it unclear how the latter could justify a judgement—think of all those cases in which odd confabulations are mentioned to back up a belief—this would also turn SMT into a form of the VIT that we have already seen reasons to reject.

[3] Note in passing that VIT also appears to imply that forgetting a bad source of information while preserving a belief enhances its justification.

These last observations create what I tend to perceive as the most serious problem for SMT, namely the fact that it moves around the epistemological issue without resolving it. This is so because it remains silent on the epistemological status of source-monitoring judgements or beliefs. If a judgement about the source, such as John's judgement that he heard a reliable historian tell him that *p,* justifies the memory judgement that *p,* then this justificatory power surely traces back to its own epistemological credentials. The judgement about the source has to be justified for it to transmit justification to the judgement that *p*. If this judgement is itself a propositional memory judgement—it has to be if it manifests preservation of information about the source rather than reconstruction—this is exactly the problem we try to resolve and about which SMT remains silent. We need to discriminate justified from unjustified source-monitoring judgements, but SMT is not up to the task.

12.2.4 A Diagnosis

We have seen that the Present Reason Theory is, at least in the forms that have been discussed, at pains to account for the justification of propositional memory judgements. It is now time to draw a general lesson from our discussion so far.

You will remember that the three theories presented above qualify as Present Reason Theories because they respectively appeal to memory impressions, beliefs on which an inference to the best explanation can be or is drawn and source-monitoring judgements to which the subject has access when he makes the propositional memory judgement. Whatever confers justification on memory judgements is to be found in the circumstances surrounding the making of these judgements. This core assumption may be described as a form of 'present-tense internalism', since it combines two claims. First, that what confers justification is subject to an accessibility requirement and, second, that this requirement is satisfied when the memory judgement is made.

Present-tense internalism about propositional memory is the source of many of the problems we considered. Adherence to this thesis explains why the MIT detaches the justification of propositional memory judgements from the reasons one *had* to judge and concentrates exclusively on present impressions. It also explains why the VIT is bound to claim that, insofar as two subjects are in a given predicament with respect to the inferences they can or do *now* make, they are equally justified. Finally, it constitutes the reason why the SMT cannot tell apart justified from unjustified source-monitoring judgements. In each case, an exclusive focus on what is accessible to the subject when he remembers engenders serious problems because the epistemological relevance of what *happened* goes unnoticed.

If this diagnosis is along the right tracks, then it creates, in my opinion, a strong case against the Present Reason Theory and in favour of the PastRT, which differs from it in directing our attention to what happened before propositional memory judgements are passed. It is to this theory that I now turn.

12.2.5 The PastRT

The PastRT can be fruitfully approached by coming back to the function of what I have called, in Sect. 1, propositional memory reports. We should observe that these reports do not simply attribute knowledge preservation.[4] They rather appear to be sensitive to a complex pattern exemplified by reasons for the relevant judgement. To pin down this pattern, let me distinguish two times of epistemic evaluation: the time *at which* one remembers and the time *from which* one remembers. Propositional memory reports seem to be correctly used when a belief has been acquired for a reason satisfying the following requirement: the reason must be apt to justify the belief at the time it was acquired as well as at the time of evaluation, i.e. the time at which the 'remember that' report is used.[5] That is:

	Belief	Justification	Constraint on justification
Time from which one remembers (t_1)	Yes	Yes	A new reason must justify a belief
From t_1 to t_e	No	No	In the absence of defeaters, the reason acquired at t_1 could justify the belief
Time of evaluation (t_e)	Yes	Yes	The reason acquired at t_1 must justify the belief at t_e

Let me call reasons that satisfy this pattern *past reasons*. These reasons provide the starting point we need to move away from the Present Reason Theory. Refining the diagnosis presented in the previous section, we can now observe that this theory rejects the claim that past reasons justify memory judgements at t_e. This is why they appeal to a justification-conferring feature accessible at that time. And, since we have seen plenty of reasons to reject this move, this provides a powerful argument in favour of the PastRT. This theory straightforwardly translates the pattern exemplified by past reasons in epistemic currency: Memory beliefs or judgements at t_e are justified by past reasons.

This constitutes a radical departure from the Present Reason Theory. The most significant aspect of PastRT is indeed constituted by the absence, at t_e, of an access requirement on past reasons that justify the judgement at that time. This arguably constitutes its central virtue, since it allows PastRT to draw epistemological distinctions that the Present Reason Theory proves unable to draw. As we have seen, it is often the case that one has at t_e past reasons for judging without these reasons being

[4] As opposed to a classical analysis defended in Landesman (1962), Malcolm (1963), Munsat (1967) and Zemach (1968).

[5] Naylor defends a similar account of propositional memory reports: '*B* remembers that p from *t* iff (1) there is a set of grounds a subset of which consists of (a) only those grounds B has at both *t* and the present for being sure that *p,* and (b) enough such grounds to make it reasonable at both *t* and the present for *B* to be sure that *p,* and (2) there is no time prior to *t* such that *B* has a set of original grounds dating from that time' (Naylor 1971, p. 33).

accessible at that time. When he passes his memory judgement, John may for instance be oblivious to the fact that his reason for judging that Dreyfus was innocent is that a reliable historian told him so. Since we quite commonly are in a situation similar to John's, requiring that justification of propositional memory judgements depend on reasons accessible at t_e would leave many of them in the sceptic's hands. PastRT avoids this kind of scepticism in claiming that past reasons are enough.[6] This is what allows it, as opposed to the Present Reason Theory, to distinguish justified from unjustified judgements among those that are accompanied by the same phenomenology and by the same possibility of explanatory inference, as well as to distinguish justified from unjustified source-monitoring judgements.

Let me emphasize two further aspects of this approach. First, according to PastRT, there must be a sort of dependence between the prior acquisition of a past reason and a judgement at t_e in order for that reason to justify this judgement. John's judgement that Dreyfus was innocent is defeasibly justified at t_e only if he so judges *because* he acquired a past reason for it at t_1. His judgement would not be justified if it did depend on bad reasons he acquired in the meantime. That is, for his judgement to be justified by the past reason, the following requirement must be met: John would not judge that p if he had not been told so by a reliable historian at t_1 and would judge differently if he was told something else by that historian. The justificatory role of past reasons hangs on such dependence.

Second, this means that PastRT is distinct from the claim that one should stop believing when one evaluates one's reasons as unsound, a claim known as the 'principle of positive undermining' (Harman 1986, p. 39). PastRT does not endorse a 'default and challenge' conception of justification (Williams 2001) with regard to propositional memory judgements: these judgements are not justified until proven guilty. Rather, PastRT appeals to dependence on past reasons so as to distinguish justified from unjustified judgements among those for which one has no accessible reason. The application of the principle of positive undermining is as a result limited to judgements that satisfy this constraint: Only these are justified until proven guilty.

12.2.6 Developing the PastRT

In the previous section, I introduced the PastRT as a radical alternative to the Present Reason Theory. Since this approach to the justification of propositional memory judgements can be developed in various ways, I now want to spend some time to explain how I think it should be developed.

To see what is at stake here, note that PastRT can put different constraints on the two times of epistemic evaluation. Thus, according to a first development of PastRT, propositional memory judgements that p at t_e are justified only if judge-

[6] PastRT is of course compatible with the idea that accessible past reasons have an epistemic impact. It only denies that this is required for justification at t_e. For one way to develop this point, see the end of Sect. 2.6 below.

ments that p at t_1 are themselves justified, i.e. *fully* exhibiting the pattern presented at the beginning of Sect. 2.5 is required. An alternative development is advocated by Lackey (2005), who maintains that the judgement occurring at t_1 need not be justified for the propositional memory judgement to be justified, i.e. *partially* exhibiting the pattern is enough for justification at t_e. Let me call these respectively *Full* and *Partial*.

> Full: A propositional memory judgement is justified at t_e if it manifests a belief acquired at t_1 for a reason in the light of which the judgement is justified *at* t_1 *and* t_e.
> Partial: A propositional memory judgement is justified at t_e if it manifests a belief acquired at t_1 for a reason in the light of which the judgement is justified *at* t_e.

I shall now, as a first step in the direction of assessing the respective merits of Full and Partial, expand my previous example. Let us suppose that John acquires the belief that Dreyfus was innocent at t_1 because Mary tells him so, but that his belief is not justified because he has bad reasons not to trust her. Let us add that, between t_1 and t_e, John forgets the source of his belief and does not acquire any new reason in its favour. Finally, suppose that, at t_e, he judges that Dreyfus was innocent because he acquired the belief at t_1. Full and Partial differ in how they assess John's memory judgement. According to the former, his judgement is not justified at t_e. But according to the latter, it is justified insofar as John has forgotten his bad reasons to distrust Mary or has in the meantime come to believe she is trustworthy.[7]

I shall now argue that Full should be preferred. To start, note that Partial goes against the following epistemological intuition. It seems that if one judges that p because one has judged that p, then the epistemological status of the former judgement depends on that of the latter. For instance, even if the relevant defeater is later defeated, it seems that John judges *for a defeated reason* given that his judging at t_e that Dreyfus was innocent depends on having so judged at t_1, a time at which the reason was defeated. He would judge *for an undefeated reason* only if his judgement were to depend on a prior judgement made at a time when the relevant reason was not defeated. I am tempted to think that this favours Full, which differs from Partial precisely in its capacity to draw such an intuitive contrast. So, let us dig deeper to see whether this intuition is sound.

To do so, let me explain why Full assesses John's judgement as unjustified. I think that this verdict is motivated by the following observation. Given his situation, John would judge that Dreyfus was innocent irrespectively of whether his reason in favour of judging so is presently defeated or not. It is for that rea-

[7] Favouring Partial, Lackey (2005) argues that memory can for that reason be a generative epistemological source: The judgement is justified because it is a memory judgement. This would mean that judgements are justified at t_e because they manifest beliefs reliably preserved from t_1 and so that preservation in itself positively contributes to justification. Yet, the fact that preservation explains why the judgement is made at t_e does not imply that it plays such a role. We should rather say that preservation allows past reasons to (potentially) justify the judgement at t_e by making it the case that this judgement depends on a belief acquired because of the past reason. For judgements are not justified simply because they manifest a preserved belief; they rather inherit their justification from the past reasons. So, even though Partial were correct, this would not support the claim that memory is a generative epistemological source.

son that Full rejects the strong form of externalism about propositional memory that is characteristic of Partial. For note that, according to Partial, what justifies these judgements is the *reliability* of the cognitive mechanism that delivers these judgements through preservation of beliefs based on the relevant kinds of reason (Lackey 2005). It is, for instance, because it manifests a reliable cognitive mechanism—the memory preservation of beliefs based on what serious informers tell us—that John's judgement is now justified (if not defeated). And the fact that Partial adopts this form of reliabilism about propositional memory will now allow us to build what I perceive as a strong argument against this approach.

To see how, note that reliabilism faces the task of specifying *psychologically realistic* and *reliable* cognitive mechanisms. So, which cognitive mechanism is at play in John's case? I think there are two possible answers to this question. One may attempt to specify the mechanism either as that consisting in (a) preserving unjustified beliefs when what defeated the reasons one had for them is itself defeated, or as that consisting in (b) preserving unjustified beliefs if one has no reason against them. And this creates a dilemma.

On the one hand, (a) is surely a reliable mechanism, but it is psychologically unrealistic. To be realistic, it would have to require that one has access to the relevant reasons in order to be in a position to discriminate actually defeated from actually undefeated reasons. Now, to add such a requirement in a defence of PastRT would offset one of its main advantages over the Present Reason Theory, since this would after all hand in to the sceptic that large bulk of propositional memory judgements for which no such discrimination is possible. Yet, in the absence of such a requirement, it is fair to say that one would still judge even if the reason were in fact defeated. The mechanism being for that reason unreliable, the judgements it gives rise to are, by the reliabilists' own lights, unjustified.

I think that we should reach the same verdict regarding (b), the mechanism consisting in preserving unjustified beliefs insofar as one has no reason against them. This mechanism is clearly psychologically realistic, but is of no avail to epistemologists of a reliabilist bent. After all, preserving beliefs that were not justified in the light of the reasons we had but which are actually not defeated is surely not a reliable mechanism of belief preservation.[8]

The fact that this dilemma does not affect Full suggests that this way of developing PastRT is preferable. And we are now in a position to realize that Partial faces this dilemma because it dispenses, contrary to Full, with a specific internalist constraint on justification. Let me explain. According to Full, the fact that a memory judgement made at t_e depends on an unjustified past judgement for which one had a reason that, given how the situation has changed, could justify it is not sufficient for that judgement to be justified. Rather, the past judgement should also have been

[8] Memory is often said to be epistemologically similar to testimony (e.g. Burge 1997). If so, then my argument has the following implication. The fact that one would not share the witness' reasons against his claim is not enough for justification. What is required is that one actually realizes that the witness' defeaters are themselves defeated. Justification transfer is sometimes blocked by irrationality (or malice) and to get rid of it requires access to an undefeated reason.

justified in the light of the reason. The justification of memory judgements requires that one has, at a time, *access* to a reason in the light of which the judgement is justified. Against this, Partial admits that these judgements can be justified even if no such time ever occurs, like when the reason is accessible but defeated at t_1 and undefeated but inaccessible at t_e.

This disagreement stems from the different ways Full and Partial conceive the *basing relation* for memory judgements and its connections with the subject's epistemic responsibility and rationality. Full claims that this relation has two parts: (a) the subject must have judged because he had access to a reason in the light of which his judgement was justified, and (b) his memory judgement must depend on his having made the past judgement in this way. Thanks to (a), considerations of epistemic responsibility and rationality at the time from which one remembers have a bearing on subsequent justification. By contrast, Partial turns (a) into: (a') the subject must have judged because he had access to a reason in the light of which his judgement would now be justified ((b) is modified accordingly). For that reason, it allocates no role to epistemic responsibility and rationality in the justification of memory judgements: These judgements can be justified despite the fact that one never judges because one has access to reasons in the light of which they are justified.

Keeping the above dilemma in mind, we can conclude that the basing relation for memory judgements must be as Full conceives it to be so as to avoid a form of epistemic luck. For a memory judgement to be justified, it must depend on a reason that justifies it. This is the case if it depends on one's having made it at t_1 in the light of an undefeated reason, but not if it depends on one's having made it at t_1 it in the light of a defeated reason. For in the latter case, as I already observed, one would still make the judgement if the reason was still defeated. This form of epistemic luck is, I think, detrimental to justification.

Let me bring this discussion to a close with a few words regarding how we should develop Full. In my opinion, we should use the requirement that memory judgements must be based, at a time, on accessible reasons apt to justify them so as to distinguish several cases of propositional memory. In the most straightforward cases, this requirement is satisfied at the time the belief is formed. In more complex cases, it is not because the reason is at that time defeated. I suggest that, in these cases, subsequent justification requires that the reason be preserved so that one is able, at a later time, to base one's judgement on it. If John's reason to believe that Dreyfus was innocent is that Mary told him so at t_1, a time at which he did not trust her, then he must remember that she told him that Dreyfus was innocent at t_2, a time at which he trusts her, in order for his judgement to be justified by that reason from that time.[9] Developed along these lines, I think that the PastRT constitutes an appealing account of justified propositional memory judgements.

[9] This means that the Source-Monitoring Theory (Sect 2.3) is true in these more complex cases. The mistake is to extend it to all cases of propositional memory.

12.3 Episodic Memory

The first part of my argument consisted in explaining why the PastRT provides a more appealing epistemology for propositional memory than the Present Reason Theory. Does this mean that we should extend this theory so as to cover episodic memory as well? The second part of my argument will consist in explaining why this is not the case. As we shall see, when we turn our attention to episodic memory, the Present Reason Theory becomes almost irresistible.

12.3.1 The Specificity of Episodic Memory

We have seen in Sect. 1 that episodic memory is about specific events in one's past life, and that there exist various attempts to distinguish it from propositional memory. As I said there, my aim here is not to assess these claims about episodic memory. Yet, the following observations prove important in the context of the present discussion.

First, and as already emphasized, some influent approaches to episodic memory claim that it differs from propositional memory simply in virtue of having a distinct subject matter. As a result, they block in my opinion the possibility of developing an epistemology specific to episodic memory. Second, these approaches to episodic memory are not sufficiently sensitive to the existence of a distinctive intentional relation to past events in our life, an intentional relation that is quite unlike what happens in propositional memory. Remembering in this way events in one's own past is, as opposed to merely remembering that these events happened, to stand in a phenomenologically rich intentional relation to these events. It is this form of remembering that tempted so many philosophers to try to understand memory in terms of memory images or, more neutrally, of memory experiences.

My interest will be in this distinctive way of remembering. I happen to think that we should draw the distinction between propositional and episodic memory by means of that which constitutes one's remembering and so account for the specificity of episodic memory in terms of these memory experiences.[10] But if you disagree, you can regard what follows as an argument for drawing a distinction between two sorts of memory in virtue of their different epistemological structures. For, independently of the verbal quarrel about what deserves to be described as episodic memory, the epistemological distinction on which I shall focus is sufficiently important to play a central role in a taxonomy of mnesic phenomena.

This being said, let me now try to answer the following two questions. First, what are these experiences that constitute this distinctive kind of memory? Second, which epistemological role do these experiences play?

[10] Hoerl (2001) and Martin (2001) rightly emphasize the central role of experiences in episodic memory.

As regards the first question, one striking fact about memory experiences is that they bear a systematic and probably irreducible similarity to perceptual experiences. Consider the following examples. Remembering a melody one once heard is phenomenologically similar to hearing it, remembering an accident phenomenologically close to seeing it, remembering the taste of a wine phenomenologically close to tasting it. These memory *experiences,* it should be emphasized, are quite unlike the memory *impressions* discussed above in connection with propositional memory. Appeal to memory impressions is, as we have seen, part of an attempt to describe the distinctive phenomenology of propositional memory, of the act of remembering that *p* and this independently of its specific content. The memory experiences we are now discussing do not exhibit this content independence. An experience similar to seeing a particular accident is constitutive of episodically remembering that specific accident and different memory experience, also exhibiting this similarity with the relevant past perceptions, are constitutive of the episodic memory of a melody or of a taste. For the purposes of the present discussion, it is enough to add that these experiences re-present items that previous experiences presented and thus inherit at least part of their intentionality from the intentionality of these previous experiences.

Let me now turn to the second question regarding the epistemological role of these experiences. When we remember by means of these experiences, we unquestionably rely on them in making past-tensed judgements and consider them as justifying us in so judging. To drive this point home, note that there is an important difference between cases where one propositionally remembers that an event occurred and cases where one has a memory experience of the event. In the second case, one can refer to one's memory experience in order to back up one's judgement, which is the case when one answers a question as to why one thinks that the event happened by saying 'because I (distinctively) remember it'.

These observations seem to me to fix a fundamental constraint regarding the epistemological role of memory experiences. Any theory that tries to dispense with these experiences is deeply revisionist about our practice of making past-tensed judgements in episodic memory. But what should we more positively say about the epistemological role played by these experiences? A straightforward answer consists in claiming that these judgements are justified if based on these experiences. An appealing way of developing this answer consists in accepting something like the following principle:

> If a mnesically appears *F* to *S,* then *S* is prima facie justified in judging that there was an a which was *F* if s/he bases her/his judgement on this experience.[11]

This principle claims that when Jim episodically remembers an accident and judges that it was *F,* what justifies his judgement is the fact it is based on his memory experience of this accident. A principle of this nature seems to be needed insofar as we want to respect the intuitive role played by memory experiences in the justification of past-tensed judgements. And, since we obviously cannot extend such a principle

[11] Chisholm (1989; Chap. 5) defends a closely related principle.

to propositional memory, this supports the idea that there is a fundamental epistemological difference between episodic and propositional memory judgements. Let me elaborate on this point.

The above principle qualifies as a *Present Reason Theory*. Since the memory experiences that justify episodic memory judgements occur at the time these judgements are passed, it subscribes to what I called in Sect. 2.4 present-tense internalism. This means that the epistemology of episodic memory judgements is, according to this principle, quite similar to that of perceptual judgements: in both cases, the judgements are justified because they are based on experiences with a specific content.

Before I consider some reasons not to endorse this principle, let me end this discussion by stressing four of its features. First, its antecedent does not distinguish veridical memory experiences from memory illusions or hallucinations. I tend to perceive this as a virtue because I think that, sometimes at least, illusory or hallucinatory experiences justify past-tensed judgements. This is the case when what appears to one as a memory experience is in fact an indistinguishable mnesic hallucination. But if you side with some recent developments of disjunctivism in this respect, you can easily modify the principle accordingly. Second, the antecedent does not refer to judgements about experiences but about the world, for the often-stressed reason that making judgements about features of our experiences is not the norm (e.g. Pollock and Cruz 1999, p. 25). In this sense at least, mnesic experiences are transparent. Third, the consequent states that if the antecedent is satisfied, then judgements are justified insofar as (a) they are formed because of the relevant memory experience and (b) they attribute to the remembered object a property this experience re-presents it as having. Both requirements are needed to distinguish judgements based on experiences from judgements that merely co-occur with them. Fourth and finally, judgements based on memory experiences are according to this principle prima facie justified, i.e. justified provided no defeater is present.

12.3.2 Some Challenges

Despite the fact that the Present Reason Theory in the form of the above principle offers an intuitively appealing epistemology of episodic memory judgements, the truth is that it faces serious challenges. In this section, I shall enquire as to whether this principle can take them up.

Audi nicely expresses the first challenge on which I want to focus in suggesting that the Present Reason Theory is wrong headed because we should not

> try to find a rich phenomenal ground for every justified memory belief. That effort should be seen as very likely to be motivated by a futile desire to understand memorial justification on the model of perceptual justification, for which there is a basis quite distinct from the experience of taking in or even carefully considering the proposition in question. (1995, p. 35)[12]

[12] A similar objection is made in Naylor (1985) and Senor (1993, pp. 456–459).

We are in a position to agree with a first reading of Audi's remark, though not with a second. If Audi's aim is to criticize the attempt to model the epistemology of propositional memory on that of perception, this is unquestionably correct. We have seen that memory impressions are quite elusive and at any rate not sufficient to justify these judgements. However, as is revealed by the context surrounding the above passage, this is not what Audi has in mind. His claim is rather that it is futile to try to find *rich* (memory experiences) as opposed to *poor* (memory impressions) phenomenal grounds for the justification of memory judgements. And our conclusions should lead us to view this as an unsatisfying starting point for an epistemology of memory. What is fruitless is not the appeal to rich phenomenal grounds in connection with episodic memory judgements, but rather the appeal to poor phenomenal grounds in connection with propositional memory judgements. After all, in the light of the way memory experiences contribute to the making of past-tensed judgements, we would misdiagnose the attempt at building the epistemology of episodic memory judgements on rich phenomenal grounds if we were to view it as revealing a futile desire to model the epistemology of memory on that of perception. Yet, can we explain why as seasoned an epistemologist as Audi misdiagnoses the situation in such a way? I believe the explanation is to be found in the assumption that the epistemology of memory is unified. This assumption explains why one may be led to think that, since rich phenomenal grounds cannot be appealed to in connection with propositional memory, poor phenomenal grounds justify all memory judgements. This is precisely the assumption I want to reject and to which I shall come back in the final section. But, except for this assumption, Audi gives no reason to think that the epistemology of episodic memory is not similar to that of perception.

The second and third challenges which I want to discuss centre around issues having to do with the epistemological dependence of memory. One might first wonder, and this constitutes the second challenge, if the claim that the epistemology of episodic memory judgements is similar to that of perceptual judgements is not doomed from the start since, unlike what happens in perception, one already knows what one remembers (Landesman 1962). Now, if this amounts to claiming that the judgements we make when episodically remembering have all already been made at the time the past experience occurred, it would not constitute a good reason to reject the principle. There are after all original episodic memory judgements, judgements one passes for the first time when one episodically remembers (Martin 1992). More charitably, we may read the challenge as claiming that memory should not be conceived as an independent, quasi-perceptual access to the past. Such a conception of memory does not subtend the principle, however. The similarity with perception is claimed to lie in the fact that different kinds of experiences play the same epistemological role in perception and in episodic memory, which does not imply that episodic memory constitutes an independent access to the past.

Still, one might nourish doubts about the principle's adequacy for a slightly different reason that constitutes the third and final challenge. For, even if the principle does not rest on the claim that memory is an independent access to the past, does it not go against the epistemic dependence of memory? Naylor nicely expresses this challenge in the following passage:

> That memory impressions cannot now give me knowledge unless one has had an original justification of a sort that could (if one had it now) now give one knowledge, is a principle whose denial would mean that our memories could sometimes be a source of knowledge in an unwelcome way. (1982, p. 435; see also Annis 1980)

Remember that we already encountered this problem in connection with propositional memory. I argued there that we should not, as opposed to the Present Reason Theory, divorce the justification of propositional memory judgements from past reasons. Does the principle commit the same mistake in failing to respect the epistemic dependence of episodic memory judgements?

It is true that some of the judgements the principle counts as justified could not have been justified in the past. This is the case for judgements based on hallucinatory memory experiences. I suggested above that some of these past-tensed judgements may be justified, and this despite the fact that there are, obviously, no past situations in which the corresponding present-tensed judgement could have been justified. With regard to these judgements, the principle indeed endorses a sort of epistemic independence, which appears to be required if one thinks that memory hallucinations can justify. But one may alternatively consider this as an argument in favour of restricting the principle to veridical memory experiences. Whatever one's verdict in this respect, I think that the admission of this sort of epistemic independence is compatible with the claim that memory is epistemologically dependent. For, according to the principle, epistemic independence takes place only when, unbeknownst to one, one is not enjoying a memory experience. This is to say that epistemic dependence is secured insofar as one is ready to conceive of memory experiences as experiences that re-present what the past experiences on which they depend presented. To justifiably judge on the basis of a memory experience that the accident was *F* implies that the judgement that it is *F* was at least prima facie justified at the time of perception. And this, we may further observe, constitutes an epistemologically significant difference with memory impressions. Appealing to the latter, I argued, implies that the justification of memory judgements is independent of one's past reasons in their favour. By contrast, memory experiences depend on 'an original justification of a sort that could (if one had it now), now give one knowledge'. They re-present this original justification and, in a sense, do not provide a new one.

12.4 Resolving the Debate

Let me conclude. In Sect. 2, we have seen that a specific kind of PastRT is best suited to deal with the justification of propositional memory judgements. These judgements are justified by past reasons, and this independently of whether these reasons are accessible at the time we remember. In Sect. 3, I argued in favour of the Present Reason Theory about episodic memory judgements. These judgements are justified by memory experiences that take place at the time of remembering.

This motivates the following resolution of the debate between the Past Reason and the Present Reason Theories. This debate is the result of a mistaken attempt at

extending one of these theories to account for the justification of a kind of memory judgement for which it is not suited. On the one hand, the Present Reason Theory is correct for episodic memory, but extending it to propositional memory judgements leads to the claim that these judgements are justified by what happens at the time we remember, for instance by memory impressions. And this creates some serious difficulties. First, not only are memory impressions often quite elusive, but they are also in serious tension with the epistemic dependence of memory and insufficient to account for the justification of propositional memory judgements. Second, attempting in this way to unify the justification of propositional and episodic memory judgements typically leads one to favour memory impressions over memory experiences, an unappealing move in the light of the rich phenomenology of episodic memory. On the other hand, the PastRT is correct for propositional memory, but extending it to episodic memory runs afoul of the ways we usually justify episodic memory judgements by downplaying the epistemological role of memory experiences.

Since these problems all derive from the attempt at providing a unified theory regarding the justification of memory judgements, my conclusion is that no such unity is to be found.[13]

Acknowledgment This chapter is the descendent of Kevin Mulligan's suggestion that I might be interested in working on memory, but I guess I am the only person still episodically remembering that event. It is in any case a modest tribute to what I learnt from him.

References

Annis DB (1980) Memory and justification. Philos Phenomenol Res 40:324–333
Audi R (1995) Memorial justification. Philos Topics 23(1):31–45
Burge T (1993) Content preservation. Philos Rev 102:457–488
Burge T (1997) Interlocution, perception, and memory. Philos Stud 86:21–47
Chisholm R (1989) Theory of knowledge. Prentice Hall, Englewood Cliffs
Dokic J (1997) Une théorie réflexive du souvenir épisodique. Dialogue 36:527–554
Ginet C (1975) Knowledge, perception and memory. Reidel, Dordrecht
Harman G (1986) Change in view. MIT, Cambridge
Hoerl C (2001) The phenomenology of episodic recall. In: Hoerl C, McCormack T (eds) Time and memory. Oxford University , New York, pp 315–335
Howe ML (2000) The fate of early memories. American Psychological Association, Washington
Lackey J (2005) Memory as a generative epistemic source. Philos Phenomenol Res 70(3):636–658
Landesman C (1962) Philosophical problems of memory. J Philos 59:57–65
Locke D (1971) Memory. Doubleday & Co, New York
Malcolm N (1963) A definition of factual memory. In: Knowledge and certainty. Prentice Hall, Englewood Cliffs
Martin MGF (1992) Perception, concepts, and memory. Philos Rev 101:745–763
Martin MGF (2001) Out of the past: episodic recall as retained acquaintance. In: Hoerl C, McCormack T (eds) Time and memory. Oxford University, New York, pp 257–284
Munsat S (1967) The concept of memory. Random House, New York

[13] My conclusion has some affinities with Burge's distinction between substantive and purely preservative memory (1993, 1997).

Naylor A (1971) B remembers p from time t. J Philos 68:29–41
Naylor A (1982) Defeasability and memory knowledge. Mind 91:432–437
Naylor A (1985) In defense of a nontraditional theory of memory. Monist 62:136–150
Owens D (1999) The authority of memory. Eur J Philos 7(3):312–329
Peacocke C (1986) Thoughts: an essay on content. Blackwell, Oxford
Perner J (2000) Memory and theory of mind. In: Tulving E, Craik F (eds) The oxford handbook of memory. Oxford University, New York, pp 297–312
Perner J, Ruffman T (1995) Episodic memory and autonoetic consciousness. J Exp Child Psychol 59:516–548
Pollock J (1974) Knowledge and justification. Princeton University, Princeton
Pollock J, Cruz J (1999) Contemporary theories of knowledge. Rowman and Littlefield, Lanham
Senor TD (1993) Internalistic foundationalism and the justification of memory belief. Synthese 94(3):453–476
Tulving E (1972) Episodic and semantic memory. In: Tulving E, Donaldson W (eds) Organization of memory. Academic, New York, pp 381–403
Tulving E (1985) Memory and consciousness. Can Psychol 26:1–12
Williams M (2001) Problems of knowledge. Oxford University , New York
Zemach E (1968) A definition of memory. Mind 77:526–536

Chapter 13
The Vocabulary of Epistemology, with Observations on Some Surprising Shortcomings of the English Language

Göran Sundholm

Abstract After some observations on the conference performance of Kevin Mulligan, the chapter notes shortcomings in the English epistemic vocabulary concerning the crucial terms *knowledge, science, evidence, certainty, proof, demonstration,* and *proposition.*

Keywords Epistemology · Evidence · Proposition · Judgement · Demonstration

"No *Englishman* ever understood the process/product distinction." Those words, spoken in an authoritative, then high-pitched, slightly nasal voice, and not without considerable malice, immediately captured my interest, since in my first non-technical philosophy paper I had been concerned to apply precisely this distinction to the notion of *construction*. The speaker was Kevin Mulligan, and the occasion was the first time we met, in a gathering of participants, during the 1989 Wittgenstein Centenary in the *Volksschule* at Kirchberg am Wechsel in Lower Austria. His words were not directed against me, but who the intended victim or victims of his words were I do not recall now. Peter Simons? Barry Smith? As likely as not, *both* probably were. On another occasion, at a later Kirchberg meeting, I was myself at the receiving end of Mulligan's strictures. When in my lecture on *Rationality and Mistakes,* I had the temerity to refer to the Duke of Wellington as an "*English* military leader", I was firmly called to order by a discrete, albeit quite audible, cough from Kevin's corner.

Since that first occasion, there have been many more meetings, sometimes at Kirchberg, where one of the more amusing exploits involved Kevin's bribery of a waiter to let him out via a back door at the *Tausendjährige Linde,* but mainly elsewhere, for instance, in Copenhagen, when we both served on a Lund University Evaluation Panel, or impromptu encounters at *Vrin's* in Paris. More often than not, good food and wine have been involved, and my children remember "Professor Mulligan" as the most convivial dinner guest in my house. Kevin likes to be kind; thus, recently at Cracow, in spite of having already dined in the most august company possible, he took me, who arrived—tired and irritable—six hours delayed, to a nice fish restaurant and made sure that I was properly fed. Twice over the past

G. Sundholm (✉)
Leiden University, Leiden, Netherland
e-mail: b.g.sundholm@hum.leidenuniv.nl

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_13,

decades, I have not taken his advice on where to eat, and both times with highly disadvantageous consequences; I have resolved it shall not happen a third time.

However, that first *obiter dictum* of Kevin's has always remained with me; it was spoken, half in jest, as part of mildly provocative banter, but nonetheless it set me thinking. Now, after more than 20 years, I would even go further than my distinguished Geneva colleague:

Not only is the process/product distinction not well understood among English epistemologists, but, by and large,

> English is not at all well tuned to the needs of epistemology.

It is my aim in this note in Kevin's honour to point out and exemplify some perhaps surprising shortcomings.

Undoubtedly, the first and foremost cause for these lies with the verb *to know*. The German, Dutch, and Scandinavian languages each have a pair of word at their disposal *kennen/wissen* (Ger), *kennen/weten* (Du), *känna/veta* (Swedish). Similarly in French, we have *connaitre* and *savoir,* and in Italian (Latin) *cognoscere* and *sapire* (*scire*), whereas in English we have only one verb *to know.* The difference between the two verbs in the pairs is that one is used for expressing "knowledge of objects", whereas the other expresses "knowledge of truths". Thus, for instance, "I know Kevin Mulligan; I know that he is a Professor." is translated into Dutch as "Ik ken Kevin Mulligan; ik weet dat hij een hoogleraar is." Knowledge of *objects* is expressed by means of *kennen,* whereas knowledge of *truths* is expressed by *weten* or *wissen*. Of course, we also have *Wissenschaft* (*vetenskap, scientia* from Latin *scire*). Science, of course, is a term that has narrowed its meaning incomparably, whereas *Wissenschaft* (*wetenschap, vetenskap*) has retained its broad meaning: "Arts and Sciences" originally meant something quite different from how it is now usually taken. English, sadly, has jettisoned the fine verb to *wit*. Today, we only come across it in the King James Bible: "God wotteth…" and in such terms as *witness, witty,* and the idiomatic use of "to wit*".* Had it still been current, *witcraft* might have served as a fine translation of the title of Kant's third *Critique*. Also, the *active* side of knowing is no longer part of the meaning of the verb, except perhaps when considering "knowledge in the biblical sense".

Most Anglophone analytic philosophers that I have spoken to, once they have understood—of course, if Mulligan was right at Kirchberg, here, perhaps, it rather ought to be: tried to, but failed to understand fully—*this* distinction, implicitly appeal to the principle of expressibility that what can be expressed in one language can also be expressed in any other language, and say that the lack of the verb *to wit* is not a serious impoverishment, since the "objects versus truth" distinction will enable one to say everything in English that can be said in German or French or Dutch or Swedish or…. That might be so, but in those languages grammar automatically points the way to, or takes care of, certain distinctions, whereas in English one has to first hit upon these distinctions and then see how they might be expressed. Clearly, it is much more difficult to draw a distinction when one has to find it *ab novo* than when it is a part of the linguistic fare and readily served up on a plate of grammar. Furthermore, with potentially disastrous consequences, the phrase "know

a proposition" becomes ambiguous between the object and truth readings, between grasping what is said, and between knowing that what is said is true.

Related to this is the treatment of *Gewissheit*. Wittgenstein's *Über Gewissheit*—a work to which Kevin Mulligan has devoted much thought, and on which, if I remember correctly, he has even supervised doctoral work—is very hard, nay, almost hopeless, to translate into English. *Certainty*, the customary rendering of *Gewissheit* in English, is also the proper translation of *Sicherheit*. Thus, one needs another English word in contrast to *certainty* that would bring out the difference between *Sicherheit* and *Gewissheit*. *Certainty,* certainly, is an adequate translation of *Sicherheit*. It captures the notion of assurance, of putting one's hand into the fire, of offering a guarantee in order to meet an obligation. When you request a loan, the building society or bank will ask for a certainty, for an assurance that the loan is safe. Here, the appropriate terms in the Germanic languages are *Sicherheit* (Ger), *zekerheid* (Du), and *säkerhet* (Sw). One could not, however, put up a *Gewissheit* as guarantee for a loan. In Dutch law, I have been told that, the testimony of a witness has to be, at least to the satisfaction of the court, "wis en zeker". We could perhaps translate this with "grounded and firmly held". The *wis* is a cognate of *weten,* the kind of knowledge pertaining to truths. Thus, the statement of a witness must be appropriately—objectively—grounded, for instance, by the witness being present, so that he could see what happened, and, secondly, it must be subjectively grounded in the sense that the witness is certain of what he saw, that his memory impressions are firm. In Swedish the same pair exists as in German: *visshet* versus *säkerhet,* where the latter has the same aspect of assurance. I have nothing to offer here as to the proper translation of Wittgenstein into English. Perhaps the best would be to use *certainty* for both notions and give the German in brackets to show what the original word is. This, after all, is the current fashion in high-minded Anglophone Frege scholarship regarding, for instance, *Begriffsschrift* (the ideography, not the book!), *Sinn,* and *Bedeutung,* in which terms are, in such works, as often as not left without translation. In this case of the Fregean *Begriffsschrift*, though, this gambit is otiose, since *ideography* is the perfect English translation as stressed by Kevin's erstwhile Geneva colleague Jonathan Barnes in a fine paper published in *Dialectica,* Vol. 56 (2002). The absence of a proper translation for *Gewissheit* is patently related to the absence of a live proper English equivalent to *wissen*.

A further very deviant usage in epistemological English concerns the fundamental notion of *evidence*. This according to the first meaning in the Oxford English Dictionary (OED) is:

> The quality or condition of being evident; clearness, evidentness.

It might come as a surprise, but this *is* the first meaning the OED offers for evidence. However, no Anglophone philosopher—not even Kevin Mulligan, as erudite an *afficonado* of Husserl and Reinach as any, who, when I drew his attention to this OED explanation of *evidence,* stated that one ought to use "self-evidence" instead—will read *evidence* as meaning the evidence *of* what is evident, but always, presumably under the influence of Anglo-Saxon legal parlance, and contemporary Anglo-American Philosophy of Science, take it as evidence *for*. James Allen's *Inference*

from Signs (Clarendon Press, Oxford, 2001) is the only recent Anglophone discussion I have found where this is even noted. The author, however, does not share my scruples, but takes quiet pride in English deviancy:

> Cicero introduced *evidentia* …the quality of being evident…In this sense it entered European languages, including English, where, however, one tends to speak of "self-evidence" because *English uniquely recognizes the sense of evidence* at issue…

Here the emphasis is mine, and the sense at issue is the British one of evidence *for* rather than *of*. Sadly, this "legal" usage is being exported back—by philosophers of science—to German and one can find examples of *Evidenz für* in current German articles in *Allgemeine Wissenschaftslehre*. However, Mulligan's (and Allen's) proposal to use *self*-evidence as a translation of, for instance, German *Evidenz* in its proper, uncontaminated sense will not do. In the case of a *propositio per se nota,* its evidence is grounded in the *propositio* itself and then *self*-evidence will be apt. When the evidence is *mediate,* rather than immediate, though, *self*-evidence clearly is out of place, since its evidence is not grounded in the claim itself, but in that of other claims. The customary use of the term *immediate* in the distinction between immediate and mediate evidence might be a further source behind the British confusion; this immediacy is in no way *temporal*. Something self-evident need not be at all *obvious* or patent. On the contrary, a mathematical axiom, taken in the old-fashioned sense of self-evident judgement, but not in the modern senses of a hypothetico-deductive assumption, or of (a component in) a structure-theoretical definition, say, an induction principle with respect to a highly complex well-founded ordering, might pose a hard challenge conceptually. Considerable experience might be needed in order to familiarize oneself with the concepts in question before the self-evident status of the axiom is grasped. It is not for nothing that the scholastics, for instance Thomas Aquinas in *Summa Theologica,* Q II, Article 1, considers propositions that are *per se nota (in se), per se nota ad nos* and *per se nota ad sapientes*. We might perhaps render this tripartite Thomistic distinction as *self-evident,* self-evident *to us,* and self-evident *to the learned.* In German *per se nota* was rendered elegantly as *an sich Erkanntes* by Horst Seidl in his translation of Aquinas' *Five Ways.*

Proof is another unhappy term that unfortunately is omnipresent in current Anglophone Philosophy of Mathematics, for instance in *proof theory,* which is the translation of German *Beweistheorie*. Dag Prawitz told me that he once received an invitation from a Swedish university to come and give a lecture on "provteori", which in Swedish means something like *the theory of tests*. What had happened was that Swedish *bevisteori* was translated into *proof theory,* which in turn was mistranslated back to Swedish as *provteori*. The event is reminiscent of the old chestnut concerning the automatic translation of: "The spirit is willing but the flesh is weak" (Matthew 26, 41) into Russian, and back again, as: "The vodka is strong, but the meat is rotten." Surely it would here be preferable and profitable to use *demonstration* instead, which is cognate with *demonstratio* (Lat), *dimonstrazione* (It.), and *démonstration* (Fr), as well as with the Germanic *Beweis*. Using *demonstration,* furthermore, has the advantage that it does have a process reading that is absent from *proof,* a point, I suspect, that might be appreciated by the honorand, in view of his

erstwhile Kirchberg claim. *Proof* has an entirely different origin and derives from *probare,* putting to the test. Accordingly, it is the *test* (but not the demonstration!) of the pudding that lies in the eating thereof. An approved man, a *vir probatus,* similarly is a tried and tested man. The rum served in the Royal Navy had to be "proof". This meant that it had survived the test of being poured over gunpowder, which then still had to ignite. *Proof* has cognates both in French (*preuve*) and Italian (*prova*). Native speakers inform me that those words are not used for the *act* of demonstration, but have only objectual uses.

Proposition, perhaps, provides the most bothersome case of all. In the tradition, a *proposition* is either a judgement (made), that is, what is "propounded", or it can also be the spoken or written garb of such a (mental) judgement, with or without assertoric force. *Enunciation,* or *sentence,* or *statement,* would do as well here, perhaps. Sometimes, in old-fashioned mathematics, a proposition can also be a *demonstrated theorem,* e.g. the Propositions of Euclid. However, a proposition can also be (the formulation of) a proposal or alternative for action. In stilted parliamentary language, "proposition has been put and a vote shall be taken". A multiple ambiguity similar to that of the old-fashioned notion of proposition is exhibited—at least from Cook Wilson onwards—by the anodyne Oxonian term *statement*. I once distinguished *seven* different uses of the term statement in one and the same Oxford exposition of Philosophical Logic. I had rather fun doing so, but the students in the third-year seminar on whom I had unknowingly inflicted the text in question, since I was unaware of its infelicities, were perhaps less amused.

Matters really became awkward in 1903 with Russell's (mis)translation of the Fregean *Gedanke* in his Frege Appendix to the *Principles of Mathematics*. There, a *Gedanke* ("Thought") is rendered *proposition*. With this, the earlier confusions become acute, since propositions have now moved from the level of *judgements* also to that of judgemental *contents*. Finally, we have also the notion of proposition used by the logician in his well-formed formulae (wff), for instance, in the propositional and predicate calculi. Following Frege's lead in the *Grundgesetze,* § 32, these formulae are associated with conditions for them to be Names of the True. In order to obviate some of the difficulties they cleave to Frege's doctrine of sentences as names of truth-values, let us say that to each wff *A* there corresponds a condition TCA such that the wff is true if that condition is fulfilled. The truth-conditions for a complex wff, say *A*&*B,* is then obtained by recursion from the truth-conditions for *A* and for *B*:

T C*A*&*B* is fulfilled iff TCA is fulfilled and TCB is fulfilled, etc.

However, at this level, no content or sense is provided. We operate solely on the level of reference, but not on that of *sense*. In order to endow the wff's with content from their truth-conditions, Frege then said that the *sense* ("Thought", *Inhalt,* content) expressed by the wff *A* is *that TCA is fulfilled*.

With this move, the resulting confusion has become irreparable. A proposition in an Englishman's text might now be a meaningful sentence, a truth-condition for such a sentence, the content that such a truth-condition if fulfilled, an assertion that

what is said in such a sentence is true, or any other out of many more or less likely options …

And on that happy note I safely deliver my observations, not into the hands of an Englishman, but into those of Kevin Mulligan, colleague and friend.

Acknowledgments At the request of its Rector, Prof. Björn Wittrock, I first presented these considerations in a lecture at *The Swedish Colloquium for Advanced Study,* Uppsala, on February 11, 2009, on the occasion of a visit from the Unit for Advanced Study, Faculty of Social Science, University of Tampere. Subsequently they were repeated in a seminar at the *Archives Poincaré,* Nancy, on April 28, 2010, during the tenure of a visiting professorship. I am indebted to both institutions for their generous hospitality and to my audiences for their helpful discussion.

Chapter 14
The Blurred Hen

Clotilde Calabi

Abstract I first present three philosophical theories on blurriness. The first theory says that seeing *x* blurrily is unlike seeing *x* as fuzzy; the second theory says that seeing *x* blurrily is seeing *x* as fuzzy; the third theory says that seeing *x* blurrily is seeing *x* without sufficient information on some of its surface visual details. I endorse the third theory. Then, I address the question whether blurriness can be considered a perceptual illusion. I argue that it can be a perceptual illusion and hence can involve some kind of perceptual error, without being a case of mismatch between perceptual content and things out there. In fact, I believe that the popular idea that illusions are mismatches between perceptual content and things out there is seriously flawed. In defending my claims, I rely on Kevin Mulligan's theory on visual awareness and primitive certainty.

Keywords Blurred vision · Fuzziness · Phenomenal properties · Visual field · Visual illusions as discrepancies between perceptual content and reality

14.1 Kevin and Sam

Kevin Mulligan certainly has many interesting ideas about seeing. Like many other philosophers, he claims that seeing, in the simplest cases, does not involve concepts and beliefs. But unlike some of these philosophers, he claims further that seeing, in these simplest cases, may have as its object states of affairs. For example, Sam may not only see a table in a room, but also see the state of affairs that consists of the fact that the table is brown. Seeing, in this latter case, is a case of visual apprehension. Visually apprehending such a state of affairs is unlike *seeing that* the table is brown, because, as we all know, *seeing that* involves beliefs and concepts. In particular, to apprehend visually the state of affairs that consists of the fact that the table is brown (if visual acquaintance is appropriate in some way) is to be acquainted with the table and its brownness.

One of Kevin's most interesting ideas about seeing in the simplest cases is that seeing involves a particular type of certainty that he refers to as "primitive visual

C. Calabi (✉)
Universita degli Studi di Milano, Milano, Italy
e-mail: clotilde.calabi@unimi.it

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_14,

certainty". Of course beliefs, too, involve some kind of certainty, but according to Kevin there is a big difference between the type of conviction characterizing critical beliefs, that is, beliefs based on cognitive activity, and his primitive certainty:

> There is a type of belief or conviction that is too primitive to be any sort of cognitive or critical belief, namely primitive certainty. Primitive certainty underlies cognitive belief, disbelief, doubt, etc. The latter typically emerges from primitive certainty. Primitive certainties are what we count on unquestionably, what we take for granted or presuppose. Primitive certainty does not admit of degrees as do the beliefs engendered by our cognitive activity. Primitive certainty has an opposite: primitive uncertainty. One is certain that *p* or not-*p*. But one is uncertain whether *p* or not-*p,* that is, simply perplexed. Belief too has an opposite, disbelief. But disbelief, like the beliefs distinguished above, is always cognitive, critical. Primitive certainty, unlike cognitive beliefs, is groundless. (Mulligan 2003, p. 35)

I find Kevin's primitive visual certainty very useful for understanding something strange that happened to Sam. I will now turn to this strange something.

Sam is mildly short sighted and wears glasses. One day he was looking at a black hen with white plumage on head and neck. The hen was five meters away from him and he could see it well. In particular, he was seeing sharply where the white plumage ended and the black started. As we might say, Sam was not seeing the boundaries between the white and black plumage as fuzzy. In fact, the boundaries of the white plumage of that hen were quite sharp.

Sadly, Sam suffers from absence epilepsy. During seizures, he blacks out and is not responsive. Each spell lasts for ten seconds and ends abruptly, and often he is not aware of anything that has happened during the spell. These episodes can occur several times each day. On one particular day, while he was looking at the hen, he had a seizure, during which his glasses fell off his nose. When he recovered, he looked at the hen again and saw it blurrily. He did not realize that his glasses were no longer on his nose, and he wondered: "This is funny. Am I misremembering? The hen looks different than it did two minutes ago. The white plumage now looks fuzzy, but I was certain it was not."

In what follows, I have two goals: first, I want to account for blurriness; and second, and more importantly, I want to understand in what way Sam went wrong. He was primitively certain about something and then a moment later he was doubtful. In fact, a mistake occurred, but what type of mistake was it? Was this a perceptual illusion, a cognitive illusion or something else?

14.2 Blurrily Seeing X and Seeing X as Fuzzy

Let me make two preliminary remarks. The first is that "blurrily" is an adverb that modifies the verb "to see" while "fuzzy" is an adjective that refers to an intentional property, that is, a property of what is seen. The second is that, as Kevin also notes, simple seeing can be understood in two versions. In one version, we directly see things in virtue of visual content and visual content is the way we see what we see. In the other version, simple seeing involves no content and the way we see what we

see is some aspect or feature of what we see. All philosophers I shall discuss adopt the first version of simple seeing.

Some of these philosophers in fact claim that blurrily seeing something is seeing it as fuzzy. They contend that when Sam saw the hen blurrily, and hence blurrily saw the boundaries of the white plumage as well, his experience was indistinguishable from the experience of seeing those boundaries as somewhat fuzzy. For others, blurriness is a property of the experience of seeing that is unlike any property of what is seen: To see something blurrily is unlike seeing *it* as fuzzy. If this were so, Sam would be able to experience the difference.

A simple idea concerning illusions remains popular among friends of visual content: The idea is that perceptual illusions are mismatches between perceptual content and things out there. Put otherwise: Illusions occur only if we perceive things as being a certain way, but they really are a different way. On this count, when Sam saw the white plumage as fuzzy, he was experiencing an illusion: The content of his perceptual experience did not exactly match the way things were. In other words, the conditions of veridicality of Sam's experience were not satisfied and he was making some kind of mistake. Kevin, who is a friend of states of affairs, could maybe assert that on that occasion Sam had a nonconceptual visual acquaintance of the state of affairs that the hen had white plumage with fuzzy boundaries.

Assuming that illusions are mismatches between perceptual content and external reality, for some philosophers blurriness is a visual illusion (there is mismatch between perceptual content and outer reality), for others it is not (mismatch of the above kind does not occurs). Those who deny that blurriness is an illusion do not generally question the simple idea about illusions. They just deny that blurriness affects perceptual content and, as a result, blurriness is not a perceptual illusion.

I disagree with all of the philosophers described thus far. Blurred vision *can* be a perceptual illusion and hence can involve some kind of perceptual error, without being a case of mismatch between perceptual content and things out there. In fact, I believe that the idea that illusions are mismatches between perceptual content and things out there is seriously flawed. In defending my claim, Kevin's primitive certainty will have a pivotal role.

In what follows, I first present three philosophical theories on blurriness and then address the illusion problem. Let me begin with the theory according to which the experience of seeing things blurrily is unlike the experience of seeing them as fuzzy.

14.3 Seeing X Blurrily is Unlike Seeing X as Fuzzy

For Kent Bach, Sam experiences the difference between blurrily seeing the hen and seeing the boundaries in its plumage as fuzzy:

> there are some phenomenal properties that really are attributable to experiences themselves…. For example, visual experiences can become blurry, as when one removes one's glasses, without their objects appearing to have become fuzzy. The objects look different, of course, but do not look to have changed. (Bach 1997, p. 467)

The question is this: What does it mean for an *x* to look different to an observer S without appearing to *S* as if *x* had changed. To answer this question, it is instructive to look at another very common phenomenon that occurs when the focus of our vision changes. Friends of phenomenal properties such as Bach contend that in this case things look different to us though the difference does not concern their visual properties, that is, properties represented in content. Again, things look different without looking as if they have changed. If things *look* different there must be some introspectible difference between the two experiences. Yet, these philosophers claim that unfocussed objects that appear blurrily are not represented as being fuzzy. Thus, blur belongs to something else.

This something else is the visual field. Paul Boghossian and David Velleman put it this way:

> [By] unfocussing your eyes you can see objects blurrily without being able to see them being blurry [alias fuzzy]. None of these experiences can be adequately described solely in terms of their intentional content. Their description requires reference to areas of colour in a visual field, areas that […] become blurry without anything's being represented to you as blurry [alias fuzzy]. (Boghossian and Velleman 1989, p. 94)

The problem concerns the notion of visual field. One could either claim that the visual field is fuzzy or that it appears as fuzzy. To the latter, one could object that only what the experience represents, appears one way or another, but the experience does not represent the visual field: it represents its objects. To the former, one could object that blurred phenomenology does not even require an unexperienced fuzzy visual field. The latter objection comes from Fred Dretske whose view on the matter we should examine more closely.

14.4 Blurrily Seeing X Is Seeing X as Fuzzy

For Dretske, experiences are representations and all representations have two aspects: vehicle and content. More precisely, representations are vehicles expressing or carrying content. If an experience is blurred, either blurriness is a property of the vehicle or a property of its content. If it concerns the vehicle, this means that the vehicle has blurry features, that is, it is fuzzy. If it is a property of the content *it is a property represented in the content*. Contrary to what Boghossian and Velleman assert, it is unlikely that blurriness is a property of the vehicle: Just as we do not need fuzzy words to express fuzzy ideas—in fact we can express a fuzzy idea with clearly printed words—and we do not need a pink representation in our head to see an object as pink, similarly, we can see blurrily without having any blur (i.e. anything with fuzzy boundaries) in our head. Those who claim that blurriness must be a property of experience are confusing the properties of a representation with the properties of what it represents. For Dretske, blurriness is a visual feature that those things we are aware may not have. To have it is to be really fuzzy. If the things one sees do not have such a feature, blurred vision is misperception, that is, an illusion (Dretske 2003).

One could rejoin that Boghossian and Velleman are not saying that to see blurrily is to have a fuzzy representation in the head: They are just saying that to see blurrily is not reducible to seeing *the perceived object* as fuzzy. This is why there is a difference in Sam's experience between seeing the hen as fuzzy and seeing it blurrily. But we are back to square one: What is this difference, if it is neither a difference in vehicle (no blur in the head) nor a difference in content?

Like Boghossian and Velleman, Dretske may be in trouble. If illusions are mismatches between appearances and reality and blurred vision is an illusion, what mismatches with what? Here is a possible answer: In blurred vision, we experience absence of many details. But it is not obvious what it is to experience the absence of something, as opposed to experiencing the presence of something. This is a tricky point that requires us to pause.

Suppose that at some point a fox had chased our hen and the hen remained with only one leg (the right one). Sam wore his glasses and, as a result, saw the hen clearly. He cried out, "Poor hen!" What did he see that made him pity that bird?

There are two possible answers: (1) Sam saw the hen and, on the basis of what he saw, he believed that the hen had only one leg (the right one); (2) Sam simply saw the absence of the left leg. The first answer involves concepts and beliefs and given my focus on simple seeing, I am not interested in it; the second does not involve concepts and beliefs. Let me focus on (2). The following further question arises: Is it possible to simply see the absence of something? What would be the difference in content between simply seeing the absence of the left leg (hence representing the absence of that leg) and seeing its presence (hence representing that leg)? Suppose now that the hen was far away from Sam and Sam did not see the *right* leg. Was he seeing the absence of it and in fact the absence of both? Finally, suppose that for various reasons, Sam saw the hen so blurrily that there was nothing he could see about either leg. Again, what is it to see an absence of so many details? These are difficult questions, and the third view on blurriness that I am about to describe bypasses all of them.

14.5 Blurrily Seeing X Is Not Seeing X Well

The third theory says that when we see things blurrily, we simply do not see them well enough to ascertain some or most of their surface details; in particular, we do not see where their boundaries and contours lie. Michael Tye defends this view (in Tye 2003). Interestingly enough he takes up something of the first view and something of the second view. Like in the first view, he wants to distinguish the experience of blurred vision from the experience of fuzziness and as in the second view, he wants to account for the difference between the two experiences as a difference in content. Let us see how this works.

In what sense is there an inherent difference in content between, on the one hand, the experience of seeing a precise thing blurrily and, on the other, the experience of seeing a fuzzy thing distinctly? Tye focuses on a special case.

In a watercolour painting executed on wet paper, the edges of the coloured shapes blur. If I view such a painting while wearing my eyeglasses, I have a clear impression of a fuzzy representation. Consider now a similar watercolour painting executed on dry paper. This image has sharp edges and viewing it without my glasses, I see it blurrily. This means that I have a blurry impression of a clear representation. There is a difference between the experience of the watercolour on wet paper and the experience of the watercolour on dry paper and observed without glasses. If I look at the fuzzy watercolour with eye glasses on,

> my visual experience represents quite precisely the fuzziness (blurriness in the text) of the edges, that is, it represents that (a) the edges definitely fall between the spatial regions *A* and *B* of the paper and (b) it is indefinite exactly where between *A* and *B* on the paper the edges fall. With the clear watercolour, seen without my eyeglasses, my visual experience is silent on the precise locus of the edges; that is, my experience represents that the edges of the coloured shapes definitely fall between *A* and *B* while failing to represent exactly where it is between *A* and *B* (that) the edges lie (Tye 2003, p. 83).

For Tye, the difference between the experiences of seeing a fuzzy thing distinctly and seeing a precise thing blurrily lies in the fact that in the former situation we have a precise representation of an object that has intrinsically vague boundaries and in the latter we have a representation of an object that does not comment on boundaries enough, or does not comment on them at all.[1]

Thus, in the third view, blurriness is *not* an illusion because illusions are misrepresentations, that is, wrong comments on what actually is out there, and blurred experiences are no comment at all, either in the positive or in the negative. The third view says that blurred vision is simply poor vision, to be accounted for in terms of lack of information rather than in terms of misinformation. In this respect, the question "what mismatches with what" does not even emerge.

The third view raises the following objection. Lack of information on relevant features of a visually presented object is not peculiar to blurred vision. For example, seeing things at a distance or through dense mist and seeing things that are partially behind other things also involves lack of information. Finally, consider the one-legged hen: Is there, leg-wise, any content difference between the visual experience of the one-legged hen and the visual experience of the hen at a distance? Both experiences are equally uninformative under that respect. Where then would their difference reside? Tye could respond that the difference always resides in the quantity and quality of information carried by the contents of these experiences. In particular, the hen seen at a distance lacks information about texture and details about boundaries, more than the one-legged hen seen in proximity.

[1] "[…] in the case of seeing sharp objects as fuzzy ['blurry' in the text], one's experience comments innacurately on boundaries. It 'says' that the boundaries themselves are fuzzy when they are not. The the case of seeing blurrily, one's visual experience does not do this. It makes no comment one where exactly the boundaries are. Here there is no inaccuracy" (Tye 2003, p. 81).

14.6 The Three Theories

Let me recapitulate. I claimed that in the third theory, given that illusions are mistaken comments on what is out there and given that blurred experiences are no comment at all, either in the positive or in the negative, blurriness is not illusory: it is simply poor vision. Tye admits that we may blurrily see a fuzzy object, but again, this would not be a case of illusion: The content would simply be less informative about the object than the perceptual context allows.[2] No error, and hence no illusion.

Despite accepting Tye's account of blurriness, I reject his idea that blurred vision is not an illusion. In fact, I think that there is room here for a special type of error and hence of illusion. As I suggested at the beginning, Sam could mistake what *really* is lack of information (his glasses having inadvertently fallen from his nose) with a change in the object (no change having in fact occurred). More precisely, he could mistake the blurriness he experiences for some degree of fuzziness in the object (recall that the hen's plumage has sharp boundaries). Here are two variations on Sam's example:

1. Nora is developing and printing a photo in her lab. She notices some change in what she sees. She is not sure whether the photo has become darker or the light has dimmed.
2. Sara has lost her balloon, which is floating away. She is not sure whether it is shrinking from deflation or moving away from her.

Is it possible for Nora and Sara to establish exactly what is happening on the basis of their visual experience? Similarly, could Sam realize that his glasses had fallen off his nose on the basis of his visual experience of the hen here and now, and no other cues? It may be that there was no experiential difference between what Sam saw when his glasses fell off his nose and what he would have seen if the object had become fuzzy. These are all situations in which the subject is so uncertain about the details that he does not know what degree of indeterminacy they might have, if any. If Sam, Nora and Sara make the same mistake, can we consider it a perceptual illusion?[3] I think that we can, but before defending my claim, let me sum up the three theories that I have presented so far.

---

[2] Of course, it is also possible to have a blurry representation of something with fuzzy edges, that is, to see blurrily a fuzzy thing. Tye remarks that the difference between this experience and seeing clearly a fuzzy thing "has to do with the degree of representational indeterminacy in the experience. If the thing we see is an image (for example a painting), in seeing the image blurrily, one's experience is less definite about boundaries and surface details than the fuzziness ['blurriness' in the text] of the image warrants. In seeing the same image clearly, one's experience accurately captures the image fuzziness ['blurriness' in the text]" (Tye 2003, p. 82).

[3] Tye remarks that "in principle an experimental setup could be devised that would leave one without any way of telling from the phenomenal character of one's experience (without any additional cues) whether one has shifted from seeing a sharp screen image through a blur to seeing a suitably blurred version of the same screen image in at least some cases" (Tye 2003, p. 82). This is precisely the situation for Sam, Nora and Sara: for them no phenomenal difference occurs (without additional cues). In fact, if there were such difference, they would not be so ambivalent between

For the first theory, blurriness is an experienced property of the visual field. For the second theory, blurriness is an experienced property of the object, hence belonging to its content. For the third theory, blurriness is lack of information. Dretske's objection to the first theory is that it does not appropriately distinguish between blurriness as a property of content and blurriness as a property of vehicle. Consider now the second theory, which is Dretske's own. I said before that if in blurred vision we experience absence of details, it is not obvious what it is to experience the absence of something, as opposed to experiencing the presence of something. Call this the absence problem. As I said, Tye has a readymade solution to it: Absence of details is just a lack of information concerning those details. In other words, absence of details is an absence of comment. This means that absence of details is not a property represented in content, as my formulation of the absence problem suggested, but rather a property of content. Finally, given Tye's acceptance of the thesis that illusions are discrepancies between the way things are and the way they are represented in experience, for him blurriness is not an illusion, since no discrepancy occurs in this case and, hence, no error.

The main difference between the second and third theory follows thus: Dretske thinks that blurred vision is always an illusion, because for him it is a mismatch between perceptual content and outer object. Using Tye's terminology, we would say that it is a mistaken or inaccurate comment on the outer object. Tye, instead, denies that blurred vision is ever an illusion. For him illusions are wrong comments, and blurred vision is not a wrong comment. For him blurriness affects content in that it is a property *of content* (content is less informative than it could or should be), but it is not a property that the content attributes to the represented object, that is, a property *represented in the content*. Given that for him illusions are inaccurate comments, if in blurred vision there is no inaccuracy, there is no illusion.

Earlier I stated that I disagree with the various philosophical opinions on this matter. My own proposal is this: Notwithstanding the fact that some illusions are mismatches between perceptual content and outer reality (which generally they are not), blurred vision may be an illusion, albeit of a different kind. I contend that my analysis explains Sam's uncertainty in a better way than the way Tye and Dretske could explain it.

14.7 Illusions

Tye notes two kinds of properties concerning content: There are properties *of* content and properties represented *in* content. On the one hand, contents can be true, false, vague, informative, not sufficiently informative, thought, believed, visually

---

the two options or even mistake the one for the other. Curiously, Tye thinks that at least in the watercolour and other similar cases, a phenomenal difference can be detected even in the absence of other cues. I do not see how the watercolour case is in any sense different from these other cases (but I do not want to question him on these grounds).

experienced, experienced in a sequence, etc. These are properties of content. On the other, contents represent properties such as being dark, being red, being five meters long, etc.

I contend that as subjects of a representational state, we can make two kinds of error concerning properties of content:

a. We can attribute to content a property that it does not have. This happens when, for example, we hold false beliefs. In this case, the content of our belief has the property of being false despite the fact that we take it as true.
b. We can mistake a property of content for a property represented in content.

Sam, Nora and Sara make the latter type of error. In particular, Sam had the problem of establishing whether the blurriness he experienced was a feature of content or represented in content (the content represents the object as fuzzy) and he got it wrong. This is precisely my point: I contend that if he made this error, he had a perceptual illusion.

My analysis requires departure from a popular account of illusions, according to which illusions are discrepancies between perceptual content and outer object. Let me focus on the weaker claim that discrepancy is a necessary condition for illusion: A perceptual experience is an illusion only if its content represents an object *O* as having property *P,* but the object *O* does not have the property *P,* an error having occurred at some point in the psychophysical chain. A number of psychologists and a few philosophers have criticized this definition on several grounds.[4] If I am correct, my blurred hen case outlined above is one more piece of evidence against it.

Those who criticize this definition question the idea that *either* the appearance (i.e. the content) mismatches with reality and hence there is illusion *or* the appearance (i.e. the content) corresponds to the way things really are and hence there is no illusion. They argue that both reality and appearances can be different things. In fact, reality can be proximal stimulation, the distal object in certain conditions, what is measured by the photometer, the meter or the scale. The same goes for appearances: There is a sense in which a circle looks elliptical from different viewpoints and there is a sense in which the circle looks as if it were a circle and not a square. It comes as no surprise then that there is wealth of counter examples to necessary conditions for illusion. Schwartz (2012) invites us to consider the Zoellner drawing, which is traditionally considered an illusion. He remarks that the parallel lines in the lower figure do not match phenomenally, and yet they are the same lines that look parallel in the upper figure. This suggests that there is a conflict between appearance and reality. The objection is that the lines in the upper figure are not exactly like the lines in the lower figure, the difference being the hatch marks. If reality includes the hatch marks, the question arises what is the conflicting appearance.

[4] For further exploration on this topic see Schwartz (2012), Maund (2012), Bruno (2012), Savardy et al. (2012).

In fact, as Schwartz notes, there is a sense in which we can place the blame on the hatch marks, because they render the two figures not exactly alike. Thus, no illusion occurs. But there is also an obvious sense in which we can claim that in experiencing the lower figure we have an illusion because both figures contain parallel lines. Colour experiences are even more instructive in this respect: If we claim that surfaces with the same reflectance spectra should match phenomenally, mismatch would mean that the experience is illusory. Given that in everyday situations the experience of colour depends on illumination, background and spatial relations, we will have the undesirable consequence that our colour experience is riddled with illusions.

It is reasonable then to abandon that definition of illusion. However, I am not suggesting that we should entirely abandon the idea that illusions are wrongful perceptions nor should we claim that illusions can never be cases of discrepancy between content and object. In fact, in cases such as the Mueller Lyer Illusion, one can still claim that the measured length of the two lines is exactly the same, yet they appear to be of different length. Thus, something wrongful has occurred. This fact, however, is not an evidence for the idea that illusions are departures from an unexperienced reality (this would require stepping outside the experience to compare it with non-experienced facts). In fact, we should seek a more profitable definition

of illusion. Given the relation between illusion and error, our definition of illusion should be broad enough to cover all perceptual situations for which we are inclined to think that an error of some kind has occurred, among which we can include the error that I have described.[5]

Let me recapitulate this. Suppose that the content of an experience represents an object with some properties but gives no information relating to other properties. There is a property of the content (the experience being uninformative on the other properties) that is not a property represented in the content. One could mistake one type of property for the other.

Finally, I should underscore that I believe that this error is likewise involved in other perceptual phenomena that are unquestionably marked as illusory. One case in point is stroboscopic movement. In fact, there is an interesting symmetry between blurred vision that occurs when the viewer changes focus (without realizing that she is doing so) and stroboscopic movement. I could even argue that the unsophisticated viewer who experiences the movement is like the person who strangely sees the objects changing when he shifts focus.[6]

Suppose that Nora at t_n sees x in l_1, that at t_{n+1} she sees y in l_2 and no longer sees x in l_1. Now, x and y are (part of) the contents of her acts of seeing respectively at t_n and at t_{n+1}. Thus, there is a sequence of two representations that Nora mistakes for the representation of the same object moving from l_1 to l_2. The illusion resides in the error of taking a property of content for a property represented in content (that is, part of the content).

The same holds true for Sam's strange experience. Again, when Sam contemplated the white plumage while under the impression that it was fuzzy, he was mistakenly taking a property *of* content (lack of information) as a property represented *in* content.

Let me now return to Kevin's primitive certainty. One objection to my account of Sam's experience is that Sam's mistake actually affected his beliefs rather than his perceptions: He ended up believing that the white plumage was fuzzy based on what he was seeing, and that belief was false. Thus, Sam's illusion was cognitive, instead of perceptual. Here is my reply. Sam's experiences might certainly have produced a wrong belief, of course. But my story is slightly more complicated. Before being

[5] Notice too that given the definition of illusion as a discrepancy, the blurred hen would be an illusion only if we endorse Dretske's account of blurriness. I have said, however, that his account requires a solution to the difficult absence problem.

[6] In considering change of focus, Tim Crane suggests that "if you didn't have the appropriate background belief you might think that you have magical powers and that the world is always bending to your intentions, becoming more or less blurred [fuzzy]." Crane's further remarks that: "It is certainly true that subjects need not to take the world to have changed, in the sense that they would judge it to have changed or believe that it has changed. But all this shows, again, is the difference between perception and judgement/belief. So removing your glasses does not change the way you would judge the world to be, in normal cases. But there is still a change in the content of the experience, in what you would put into words. You might say 'things look blurry now, even though I know they are not.' And it makes sense to suppose that someone might come to believe, because of some strange background belief, that things were actually that way […]. There is, then, change in the intentional properties of the experience, despite the fact that normal subjects would not judge the world to have changed" (Crane 2001, pp. 143–144).

stricken by the spell, Sam was primitively uncertain about the scene he was seeing, and while recovering started to nourish primitive certainty that the hen had fuzzy boundaries. In becoming primitively certain of this new state of affairs, he took as a property of the hen, that is, a property represented in its content, what was in fact a property of the content of his experience.

Acknowledgments I presented versions of this chapter at workshops in the philosophy departments of Parma and Bergamo. Thanks to the audiences in these workshops and particularly to Andrea Bianchi, Bill Brewer, Tim Crane, Jerome Dokic, Wolfgang Huemer, David Hughes and Alberto Voltolini. Special thanks to Marco Santambrogio.

References

Bach K (1997) Engineering the mind. Review of Dretske's naturalizing the mind. Philos Phenomen Res 58:459–468

Boghossian P, Velleman JD (1989) Colour as a secondary quality. Mind 98:81–103

Bruno N (2012) Illusions that we should have, (but don't). In: Calabi C (ed) Perceptual illusions: philosophical and psychological essays. Palgrave McMillan, Basingstoke

Crane T (2001) Elements of mind. Oxford University Press, Oxford

Dretske F (2003) Experience as representation. Phil Issues 13:67–82

Maund B (2012) Perceptual constancies: illusions and veridicality. In: Calabi C (ed) Perceptual illusions: philosophical and psychological essays. Palgrave McMillan, Basingstoke

Mulligan K (2003) Seeing, certainty and apprehension. In: Fossheim H, Mandt Larsen T, Sageng JR (eds) Non-conceptual aspects of experience. Proceedings of the 2000 Melbu Conference on Non-conceptual Content. Unipub Verlag, Oslo

Savardi U, Kubovy M, Bianchi I. (2012) The genesis of the awareness of illusions. In: Calabi C (ed) Perceptual illusions. Philosophical and Psychological Essays. Palgrave McMillan, Basingstoke

Schwartz R (2012) The illusion of visual illusions. In: Calabi C (ed) Perceptual illusions: philosophical and psychological essays. Palgrave McMillan, Basingstoke

Tye M (2003) Consciousness, color and content. MIT Press, Cambridge

Chapter 15
How Picture Perception Defies Cognitive Impenetrability

Alberto Voltolini

Abstract According to the thesis of the cognitive impenetrability of perception to thought—from now onwards, (TCI)—both the phenomenal character and the intentional content of perceptual states are impermeable to states of their subjects' cognitive systems. This means that no change in the content of the latter states alters either feature of the former states. Now, perception of ambiguous figures is held to be a prima facie counterexample to (TCI), for what one takes to be the picture a picture *of* influences what experience she has when facing the picture, hence it induces two different (picture) perceptions. A defender of (TCI) may well reply that in the Gestalt switch involving ambiguous figures there is indeed a phenomenological change, yet this change is only indirectly driven by the states of the cognitive system involved. For, first, those states rather induce a shift in attention, and second, this shift of attention is responsible for the phenomenological switch. Yet let us consider first the fact that in perception of ambiguous figures attention works differently than in the ordinary perceptual cases in which there is no real cognitive penetration; namely, as an active focusing on the very same elements of the figure to be alternatively grasped rather than as a focusing on a different part of the scene one was previously facing. Moreover, let us take into account the fact that picture perception of ambiguous figures is just a borderline case of ordinary picture perception, for picture perceptions both of ambiguous figures and of 'normal' figures are characterized by the lighting up of aspects (different aspects in the former case, just one aspect in the latter case). Then, the above reply may be appropriately circumvented. Insofar as in picture perception attention performs a grouping job of the very same elements of the figure one is facing and such an attentive job may suit a conceptual research, concepts mobilized by the states of the cognitive system involved may well help attention to perform such a job by conceptually informing the picture perception a subject entertains.

Keywords Cognitive (im)penetrability · Picture perception · Ambiguous figures · Illusory seeing-as · Organizational seeing-as

A. Voltolini (✉)
University of Turin, Turin, Italy
e-mail: alberto.voltolini@unito.it

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_15,

15.1 How Picture Perception Defies Cognitive Impenetrability

According to the thesis of cognitive impenetrability of perception[1] to thought—from now onwards, (TCI)—both the phenomenal character and the intentional content of perceptual states are impermeable to states of their subjects' cognitive systems.[2] This means that no change in the content of the latter states alters either feature of the former states. Perceptual illusions are paradigmatic examples of one such impenetrability. In the case of the Müller-Lyer illusion, although we believe (for independent reasons) that the two lines ending in oppositely oriented wedges have the same length, we are forced to see them as being different in length. As Fodor (1983) holds, perception is *modular*.

Now, perception of ambiguous figures is held to be a *prima facie* counterexample to (TCI). For what one takes to be the picture a picture *of* influences what experience one has when facing the picture, hence it induces two different (picture) perceptions. For instance, in the well-known example of the duck–rabbit figure, once one manages to see the figure as (a picture of) a *rabbit,* one no longer sees it as (a picture of) a *duck* and *vice versa*.[3] Undoubtedly, the two 'seeing-as'- experiences are phenomenologically different; the Gestalt switch is indeed a change in experience.[4] Thus, the perceptual situation at stake is utterly unlike an apparently analogous situation in which by mobilizing different concepts we limit ourselves to describe differently what remains one and the same 'seeing-as' experience with no undergoing switches, as in this case presented by Wittgenstein:

> Take as an example the aspects of a triangle [….] This triangle can be seen as a triangular hole, as a solid, as a geometrical drawing; as standing on its base, as hanging from its apex; as a mountain, as a wedge, as an arrow or pointer, as an overturned object which is meant, for example, to stand on the shorter side of the right angle, as a half parallelogram, and as various other things. (2009[4], II xi, § 162)

[1] In point of fact, I will focus here merely on *visual* perception, although the phenomena I will deal with may be found also in other sensory modalities, at least as far as picture perception also occurs in such modalities (cf. e.g. auditory picture perception or tactile picture perception). This is clearly enough for my purposes, for visual cases of cognitive penetrability are sufficient to undermine (TCI) in its generality.

[2] As intentionalists or representationalists claim, the phenomenal character of a perceptual state amounts to, or at least supervenes on, the intentional content of that state. I think this claim is wrong, yet for the purposes of this chapter I will remain neutral on it. Cf. my Voltolini (2013).

[3] Wollheim (1980[2], p. 220) holds that a way of describing things that appeals to *pictures* as figuring in the content of what a certain subject sees the relevant figure as is better than merely saying that the subject sees the figure either as a duck or as a rabbit. Yet the best description of the situation at stake would be to say that that subject sees the figure either as a duck or a rabbit in virtue of literally seeing the figure itself. For this is what seeing a duck or a rabbit *in* that figure really amounts to (for this account of the whole twofold experience of seeing-in (on which, see soon later in the text), cf. Levinson (1998) and my Voltolini (2012b), where I also apply it to the case of ambiguous figures). Once the subject further interprets the figure either as a picture *of* a duck or as a picture *of* a rabbit, then the description in question, namely that such a subject sees the figure either as a picture of a duck or as a picture of a rabbit, becomes fully legitimate. On this cf. again Voltolini (2012b).

[4] As Macpherson (2006) forcefully claims.

Let me describe the situation at stake in greater detail. To begin with, the phenomenological difference of the two 'seeing-as' experiences one has when facing the duck–rabbit figure is well matched by a difference in content between those experiences. In this respect, if by following Wollheim (1980[2]) we accept that picture perception amounts to the twofold 'seeing-in' experience of nonliterally seeing the depicted object in virtue of literally seeing a material object (a canvas, etc.), we can describe the above Gestalt switch as the transition from a certain twofold 'seeing-in' experience of 'seeing' a duck in a figure in virtue of seeing that figure to another twofold 'seeing-in' experience of 'seeing' a rabbit in that figure in virtue of again seeing that figure. Now, what prompts that phenomenological difference is precisely a change in the concepts that the two experiences mobilize—*being a duck, being a rabbit*. To put it differently, a relevant difference in how those concepts are to be instantiated manages to do what the analogous difference between the concepts recalled by Wittgenstein in the above example of the aspects of a triangle fails to perform; namely, a phenomenological change in the experience. In other terms, taking the duck–rabbit figure as (a picture of) a rabbit rather than as (a picture of) a duck involves a phenomenological difference in one's perception, while taking the triangular figure, for example as a (picture of) a mountain rather than as (a picture of) an arrow does not involve such a change. Now, the concepts mobilized in the above Gestalt switch also constitute the different content of different states of the cognitive system of the subject involved; that subject wants to see the figure either as (a picture of a) duck or as (a picture of) a rabbit. Since her later (picture) perceptions precisely make those intentions fulfilled, one may thus well say that her perceptions have been penetrated by those states. Or so I claim.

To be sure, a defender of (TCI) may immediately reply that in the above Gestalt switch there is indeed a phenomenological change, yet this change is only indirectly driven by the states of the cognitive system involved. For, first, those states rather induce a shift in attention, and second, this shift of attention is responsible for the phenomenological switch by letting the perceptual module do its—admittedly different—perceptual job.[5] According to one of the main defender of (TCI), namely Pylyshyn (2003), this is precisely how things work. For Pylyshyn, what happens when concepts mobilized by cognitive states seem to be involved in perception, or better in *early vision,* that part of perception which according to him is cognitively impenetrable, is the following. In a pre-perceptual stage, concepts trigger attention. Once attention focalizes a certain portion of a perceived scene, early vision picks up the objects, which are visually given in that portion in a completely nonconceptual way.[6]

[5] For this way of putting the reply, cf. Macpherson (2012).

[6] Cf. Pylyshyn (2003, pp. 62–63, 80–82, 86). According to Raftopoulos (2009, 2011) this is the job typically performed by *spatial* attention, i.e. attention focusing on certain locations. Theoretically speaking, another explanation is open to Pylyshyn, namely to claim, as he also says (cf. 2003, p. 64), that cognitively induced attention here operates at a post-perceptual stage, namely after the perceptual module does its job of picking up perceptually available objects nonconceptually. Still in Raftopoulos' (2009) account, this is the way *object-centred* attention works, by allowing the proto-objects early vision grasps to be apprehended as proper objects. At first blush, this is

Now definitely, attention is involved in having the two experiences, the 'duck' experience and the 'rabbit' experience. As Chisholm (1993) originally noted, once one focuses on a certain spot on the left-hand side of the figure and then lets her gaze go rightwards, she becomes able to see the duck in the figure, yet once one focuses on certain lines on the right-hand side of the figure and then lets her gaze go leftwards, she becomes able to see the rabbit in the figure.[7]

Yet first of all notice that the way focusing of attention works here is not the way it normally works in cases in which attention is conceptually driven yet no cognitive penetration occurs, as when one turns her eyes on different *parts* of the perceived scene in order to note something of a kind *K*—say, a tree—one did not note before. One may well say that in these cases, although they trigger attention, concepts constituting cognitive states do not play any perceptual role. For once conceptually driven attention lets the eyes focalize the right part of the scene, then the perceptual module does all there is perceptually to do, namely, it enables the relevant subject to see the objects located in that part of the scene by nonconceptually individuating them. Yet in the case of a Gestalt switch, one still keeps her eyes on the very same part of the figure, by simply selecting a different orientation—in our case, via a left to right versus a right to left perspective—according to which the very same elements of the figure are to be differently grasped. So, one may well guess that the work concepts perform here does not limit itself in activating attention, but it goes through such an activation in order to reach perception and thereby influence it.

On behalf of (TCI), one might reply that in this concern one can still distinguish between a *exogenous* and an *endogenous* form of attention, the former typically prompted by salient elements in the perceived scene and the latter prompted by cognitive states of the system involved. This shows that attention performs the relevant grouping—say, the 'duck' grouping or the 'rabbit' grouping—independently of any conceptual mobilization, in the sense that such a mobilization merely triggers attention in its doing the grouping job. So, if there is in the case at stake a cognitive penetration, this is at most indirect.[8]

Yet speaking here of merely indirect cognitive penetration is inappropriate, for it is inappropriate to talk as if there were two ways for the involved attention to work, one which immediately follows some scene's elements (or some further factors that

what Pylyshyn should maintain, since he holds that a conceptually penetrable perception amounts precisely to seeing-as perception (cf. 2003, pp. 51–52). Yet this explanation cannot be legitimately invoked here. For, as we will immediately see, in Gestalt groupings attention clearly plays a genuinely *perceptual* role in arranging in a certain way the elements of a scene.

[7] As Nanay (2011, pp. 558–559) underlines, in order for attention to perform that job, once eye's fixation on a certain point is given no further eye movement is required. For attention works *holistically*.

[8] Cf. Raftopoulos (2011, pp. 498–502). The attention at stake here is what in (2009) Raftopoulos labels spatial attention (see fn. 6). In (2009, chap. 2), Raftopoulos had already admitted a cognitive role to spatial attention while allowing to it not only a pre-perceptual job *à la* Pylyshyn, but also an interperceptual job. Yet, he had still said there that cognitive role is merely indirect. For although such a kind of attention enhances certain stimuli and inhibits others, the percept it applies to remains the same. In (2011) he seems to allow attention a genuinely perceptual way of operating, yet this way remains not cognitively shaped.

scene involves)[9] and another one which is merely stimulated by conceptual mobilization. Quite on the contrary, grouping operations involve for attention only one way of working, which precisely consists in seeing the scene's elements along a certain orientation, hence in a certain arrangement, rather than along another orientation, hence in another arrangement. Now, insofar as concepts are mobilized, they shape attention in precisely performing such visual arrangements. Put alternatively, in the cases at stake attention does not trigger perception as a sort of lighting enabling perception to grasp what is already out there (as merely stimulated by the outer elements).[10] Rather, attention performs a perceptual job that enables vision to rearrange the scene one is facing, by mobilizing properties that belong to the perceived element yet also depend on the way such elements are oriented. So in the duck–rabbit case, focusing on the left-hand side of the figure would not let one grasp the 'rabbit'-aspect of the figure unless one holistically rearranges in a certain way all the parts of the figure starting from its left-hand side (*mutatis mutandis*, the same holds of the opposite focusing concerning the 'duck'-aspect). (*Pace* Raftopoulos 2009, pp. 280–284; Orlandi 2011, pp. 317–318)

Be that as it may, Pylyshyn does not actually appeal to the above reply. He limits himself to saying that attention may perform different jobs; sometimes it enables the perceptual module to pick up an object in the perceived scene by letting the subject's eyes move to it, some other times it enables the perceptual module to pick up an object in the perceived scene by giving a direction, an orientation, to certain elements in that scene without the subject's eyes moving in any relevant sense (Cf. Pylyshyn 2003, p. 160, 168).

Yet Pylyshyn does not seem to notice that in the latter case attention plays a perceptually active role. For by enabling a different grouping of the scene's elements, it makes one's perception of that scene different. Thus, insofar as that perceptually active attention is conceptually driven, concepts inform (picture) perception.

I say 'Pylyshyn does not seem to notice' for in point of fact he endorses Peterson's et al. (1992) distinction between *reference-frame realignments*, what happens in the duck–rabbit case, and *(part-based) reconstruals,* where simply the parts of an experienced scene take different meanings, as in the afore-mentioned case of the aspects of the triangular figure.[11] Now, the difference between the two cases is precisely the one I have already pointed out. In the first case, determinate concepts play a perceptual role by inducing a Gestalt switch. Yet in the latter case they do not play that role, for no such switch occurs.[12]

[9] For, as Nanay (2011, p. 560) admits, in the case of a Gestalt switch at stake the scene's salient elements remain the same.

[10] As Raftopoulos often claims (Cf. e.g. 2009, p. 296).

[11] For other examples of this situation, see the cases of visual puns labelled by R. Price 'droodles' that Pylyshyn himself quotes (cf. 2003, pp. 43–44).

[12] Here I agree with Raftopoulos (2011, pp. 506–507). Not accidentally, moreover, Pylyshyn also espouses Peterson's et al. conclusion that, unlike realignments, reconstruals normally occur in mental imagery rather than in perception. For according to him in the imagery case there is no real image whose elements one can visually realign (Cf. 2003, pp. 347–349).

Moreover, the guess that concepts influence (picture) perception via the way in which they make attention rearrange a perceived scene in the case of Gestalt switches is further corroborated once one notices that all pictures are potentially ambiguous (as Gombrich 1960, p. 249 originally envisaged), that is, are such that utterly different objects can be seen in them. For example, what appear to be mere shadows on the depicted face of G.W. Bush in a portrait of the former US president turn out to be (pictures of) dark naked bodies scattered all around a lighter multicoloured surface (cf. http://hypehaus.com/artists/blogartists/jonathan-yeo1.jpg). On the basis of that reflection, it will turn out that what in the case of a 'normal' figure happens only once happens at least twice in the case of a Gestalt switch, unless further potentialities in the representational power of such a figure are discovered. More precisely, what happens twice in the case of a Gestalt switch is simply what actually happens once in the case of a 'normal' figure, namely, the *lighting up* of an aspect (Cf. Wittgenstein 2009[4], II xi, §§ 118, 140). But if this is the case, then in both cases the attentional focusing should be better described as a way of keeping oneself concentrated on one and the same scene unless something perceptually appears *in* that scene. Yet this shows that the job concepts here do again mobilizes attention in order to influence perception. For concepts orient a subject's (picture) perception in grasping what is there to be grasped over and above other properties of the material object that subject directly faces.

Consider for instance the well-known case of R.C. James' picture of a dalmatian. Here no Gestalt switch is actually involved. Yet for a subject seeing that picture may well amount for a long while to simply seeing black and white patches chaotically scattered all around within a certain surface, until at a certain moment a certain aspect lights up and that subject sees a dalmatian in the picture. In such a case, again, attention is surely involved by mobilizing the concept of *being a dalmatian*: if one is said to see a dalmatian in the figure, this will surely prompt an attentional research concerning the figure itself. Yet attention is here involved not in order to let the subject move her eyes towards somewhere else in the whole perceived scene. Rather, attention prompts the subject's eyes to go on seeing the very same (part of the) scene she is facing, until she gathers differently the very same elements of the scene she was looking at before. By grasping such elements in a new order, she comes to see what she previously failed to see; namely, the figure as (a picture of) a dalmatian, or, which is the same, a dalmatian in the figure.

That in such a case a phenomenological change occurs between the 'before' and the 'after' experience can hardly be denied. In this situation, a new complex twofold experience takes place that alters the original experience of what now is only a 'fold' of that complex experience, i.e. the literal seeing of the material object one is facing. In Lopes' own words:

> The features we see a picture surface as having may depend in part on what we see in the picture. Thus seeing a dog in [the picture of a dalmatian] causes one to see part of the picture surface as having a bounded subjective contour, which is invisible when no dog is seen in the picture. This phenomenon is widespread in pictures and affects the perceived relative size, shape, color, and contrast of part of picture surfaces. (1996, pp. 167–168)

Moreover, if that phenomenal change occurs, a change in the content of the experiences involved in that change occurs as well. All in all, we have a change from a literal seeing of a piece of paper to a nonliteral seeing of a dalmatian in virtue of literally seeing that piece of paper (Cf. Hopkins 1998, pp. 15–16). But as I have shown above, at least in ordinary cases what makes the phenomenological, as well as the intentionality, change in the subject's perceptual state is the mobilization of the concept of *being a dalmatian,* which also occurs in the content of states of the cognitive system of the subject involved. In the same vein, one may say that in the case of a Gestalt switch the conceptual difference prompts a difference in the intentional content of the relevant experiences over and above their phenomenal difference; in virtue of seeing a certain figure one now sees that very figure as a duck, then as a rabbit. Yet also independently of this matter of content, what we have in the 'lighting up' case is precisely what in the case of an actually ambiguous figure happens twice; namely, the (picture) perception's being cognitively penetrated by the relevant concept of the relevant cognitive state. Remember indeed that in order for (TCI) to be ruled out, it is enough that the phenomenal character of an experience is conceptually determined. For, as I said at the very beginning, (TCI) maintains that neither the intentional content *nor* the phenomenological character of a perceptual state is cognitively penetrable.[13]

At this point, a defender of (TCI) might retort that, although the phenomenological change in question can hardly be denied, it does not yet depend on the mobilization of a concept figuring in the content of some cognitive state. Rather, the change happens independently of the possession of such a concept.

Yet first of all, in order to see that concepts play a role in picture perception, let me evaluate once again the already recalled fact that not all concepts can induce a Gestalt switch, but only those concepts whose instantiations mobilize features that are sufficiently different. Now, let me ask the following question: In what sense exactly have those instantiations to mobilize features that are sufficiently different? Well, in the sense that the relevant perceptions involve different kinds of grouping operations that are performed at least with respect to the depicted instances of such concepts. Let us go back to the triangle case pointed out before. Why does seeing that triangle as (a picture of) a mountain rather than as (a picture of) an arrow involve no phenomenological change in its perception? For the way in which the triangle's elements are grouped remains exactly the same in both cases. Whereas, in the case of the duck–rabbit figure, the switch is phenomenologically relevant precisely because the two concepts involved—*being a duck, being a rabbit*— induce different grouping operations on the same perceived elements of the figure—as I said, a certain '*A* to *B*' grouping (seen from a left-to-right perspective) rather than a certain '*B* to *A*' grouping (seen from a right-to-left perspective): The very same

[13] According to Tye (1995, p. 140), the fact that a phenomenological change driven by certain concepts occurs in the relevant picture perception does not entail that such concepts enter into the content of that perception. This is questionable, for a subject is confronting herself with a generic picture, insofar as she sees *some F or other* in the figure. Yet independently of this, for the reasons just given in the text the fact that one such change is conceptually driven is enough for ruling out (TCI).

elements in the figure are ordered differently so that they form different perceptual unities. Suppose instead that a subject started to see the figure as a (picture of a) rabbit and then she took it as a (picture of a) hare. In that case, no phenomenological change would take place. For the concept of *being a hare* precisely induces the very same grouping of the elements of the duck–rabbit figure that the concept of *being a rabbit* induces.

Now, one way of reading the morale of the above reflection is that one and the same Gestalt switch, or even one and the same Gestalt grouping, may be induced not only by a certain concept—say, the concept of *being a rabbit*—but also by any other concept whose mobilization prompts the very same perceptual (re)arrangement of the perceived scene—say, the concept of *being a hare*. If this is the case, then the fact that a subject has no mastery of a particular concept does not yet mean that there is no cognitive penetrability of picture perception. For the relevant Gestalt grouping may be induced by another concept that subject possesses.

Moreover, I can well acknowledge that there are cases of Gestalt switches that take place spontaneously, that is, without the previous mobilization of any concept; although the description of the switch certainly mobilizes different concepts, the switch may occur even in subjects that have no mastery of such concepts. These cases involve what Wittgenstein once labeled 'purely optical' aspects (1980, I, § 1017). Purely optical aspects may be qualified as cases in which a perceptual shift may well happen independently of the fact that the figure involved has a certain representational power, is a picture *of* something; that is, the switch may well be either an alternation in seeing the figure either as (a *picture of*) a *F* or as (a *picture of*) a *G,* or an alternation in seeing the figure either as a two-dimensional *F* or as a two-dimensional *G*. Consider the 'double cross' case. One may well pass from seeing a certain figure as (a picture of) a white cross on a black background to seeing that figure as (a picture of) a black cross on a white background. But one may well merely pass from seeing a two-dimensional black figure (call it a two-dimensional black cross if you like) flanked by a white two-dimensional array of triangles to seeing a two-dimensional white figure (call it a two-dimensional white cross if you like) flanked by a black two-dimensional array of triangles. One may thus guess that in order to notice one such Gestalt switch, a subject does not need to mobilize the concepts of *being a black cross* or of *being a white cross,* as figuring in particular in the content of a cognitive state such as that one wants to see the figure as (a picture of) a white cross on a black background or in the content of a cognitive state such as that as (a picture of) a black cross on a white background. As Wittgenstein says:

> Those two aspects of the double cross (I shall call them *A* aspects) might be reported simply by pointing alternately to a free-standing white and a free-standing black cross.
> Indeed, one could imagine this as a primitive reaction in a child, even before he could talk.
> […]
> The *A* aspects are not essentially three-dimensional; a black cross on a white ground is not essentially a cross with a white surface in the background. One could teach someone the idea of the black cross on a ground of different colour without showing him anything other than crosses painted on sheets of paper. Here the 'background' is simply the surrounding of the cross. (2009[4], II xi, §§ 215, 218)[14]

[14] For other examples of merely two-dimensional switches, such as seeing a figure either as a square or a as a regular diamond, cf. e.g. Peacocke (1983, pp. 24–25), Macpherson (2006, pp. 87–90). To

To be sure, in order to see that concepts may not be involved in a Gestalt grouping of a certain figure, one does not even have to rely on cases involving merely two-dimensional switches. Going further in this direction, provided that there are no substantial differences between the case of an actually ambiguous figure and the case of a merely potentially ambiguous figure, it may indeed be the case that even in the case of the picture of a dalmatian, the only aspect that there is lights up onto a subject without that such a subject recognizes the figure as (a picture of) a *dalmatian*. Consider puzzle pictures, in which one is invited to connect points on a surface until a certain pattern lights up. Definitely, in order for that pattern to light up, no concept has to be mobilized in advance. Granted, if one is asked to say *what* she sees, she will provide an answer only if she has mastery of the relevant concept. Yet one may reasonably say that such a subject may grasp the new perceptual pattern even if she lacks such a mastery.[15]

Thus, both in the case of a figure involving a Gestalt switch and in the case of the actually non-ambiguous figure, the aspects involved may precisely merely light up onto someone rather than dawning up to her in virtue of a certain cognitive process. Yet this does not really undermine the challenge against (TCI). For remember that in order to defy (TCI) it is enough to show that perception *can* be cognitively penetrated, not that it *must* be. This is to say, what the counterexamples to (TCI) I have here presented really show is that concepts *help* one to perform in her perception a grouping operation, an assemblage of the elements perceived under a certain orientation, that might be performed even nonconceptually. The perceptual way of grouping elements attention performs may be done nonconceptually, as it may happen with puzzle pictures. This is the right way to understand the afore-mentioned distinction between exogenous and endogenous attention: one and the same kind of attention performing the very same active perceptual operation, a grouping operation, in the latter case via conceptual mobilization, in the former case without such a mobilization.[16] Yet mastery of concepts facilitates the attention necessary to do that grouping job. If there were no cognitive penetration, attention should operate such a grouping only on a nonconceptual basis. But in the examples I pointed out it is precisely the opposite that happens.[17]

---

be sure, it may well be that even such groupings involve phenomenally relevant switches that are induced by merely geometrical concepts.

[15] To be sure, a defender of the cognitive penetrability of picture perception may still retort that even in such cases concepts are hardly avoidable in order for the perceiver to perform a phenomenological switch. As Wittgenstein puts it: I suddenly see the solution of a puzzle-picture. Where there were previously branches, now there is *a human figure*. [my italics]. My visual impression has changed, and now I recognize that it has not only shape and colour, but also a quite particular "organization" (Wittgenstein 2009[4], II xi, § 131).

[16] The former case is sometimes described as if it were a pre-attentive, automatic way of performing grouping operations, for instance as regards many cases of figure-ground segmentation, in which a subject groups elements of a perceived scene in a foreground/background order (cf. Wolfe et al. 2002). But it would be better to allow that such operations, if there are any, involve attention without awareness. For the possibility of unaware attention cf. Lamme (2003).

[17] Incidentally, note that I am focusing here on the issue of cognitive penetrability of picture perception. For, as I already remarked (cf. fn. 13), if I focused directly on the issue of whether picture

To be sure, since there is no substantial difference between the case of an actually ambiguous figure and the case of a merely potentially ambiguous figure, we can well expect that even in the latter case normally concepts help one to perform the relevant grouping operation precisely in the same way as in the former case. So, even if we acknowledge that in the latter case there are situations in which the grouping operation may occur without any concept being mobilized, there are other situations in which such a mobilization is relevant: One manages to see a figure as (a picture of) an *F* insofar as she tries successfully to see that figure as (a picture of) an *F*.

Consider the well-known case of the Holy Shroud of Turin. For my present purposes, let me put aside the problem of whether that shroud is a *transparent* image of Christ—in Walton's (1984) sense of a natural sign of its meaning, Christ in this case—that really wrapped Christ's body; let me rather take it just as an image of a human body. Now, if *ex hypothesi* one were faced with that tissue without knowing anything about its interpretation, she would probably just see a bundle of dark patches scattered around it. But suppose one knew that what she has to see in that tissue is a human body. Then she would focus her attention on the very same bundle in order to let the grouping of that bundle emerge that would allow her to see it as (a picture of) such a body.

Finally in this concern, let me remark that a certain feature of the 'seeing-as' experience involved in picture perception does not have to lead one astray. It is often stressed, primarily by Wittgenstein himself, that this experience is subject to the will (Cf. Wittgenstein 2009[4], II xi, § 256). In point of fact, I recalled before that this is normally taken to mean that a subject wants to see a certain figure as a picture of an *F*. This might well lead a defender of (TCI) to think that concepts are mobilized in a sort of arbitrary operation that does not actually involve any genuinely perceptual element, so that perception comes before such an operation occurs.[18]

Yet this is precisely what is not the case. For one cannot see *at will* a figure as a picture of a *F*. Put alternatively, one may try to see something as a picture of an *F* and fail to see it as such; for instance, one cannot see a hole as (a picture of a solid with) a rectangular face (Cf. Wittgenstein 1980, II § 545). For the figure involved does not allow that seeing-as to obtain insofar as its elements cannot be appropriately grouped. Rather, in such cases will operate on attention as a means that allows a certain picture perception—rather than another one—to emerge, because of the mobilization of the relevant concept. Put alternatively, seeing a figure as (a picture of) something may well be an order to be fulfilled by means of the relevant attentional act (Cf. again Wittgenstein 2009[4], II xi, § 256). To repeat, such an act operates at a perceptual, not a post-perceptual stage. For that operation involves a grouping

perception has a conceptual content, it would be hard to escape a positive conclusion on that matter. In all the above cases, the twofold 'seeing-in' experience that features a picture perception is a *generic* one: in virtue of literally seeing certain patches of colour, one nonliterally sees not a particular object (say, my favourite dalmatian Pongo), but an object of *some kind or other*, an object falling under such kind. How can we thus avoid the conclusion that the content of a picture perception is concept-involving?

[18] This is precisely what Pylyshyn would consider a post-perceptual operation of attention. cf. fn. 6.

arrangement of the elements of the perceived scene that has perceptual import, as the change in phenomenal character in a Gestalt switch or in a mere 'lighting up' case testifies.[19]

Two final remarks before concluding. First, as Wittgenstein himself originally remarked, there are different kinds, if not notions, of seeing-as (Cf. Wittgenstein 2009[4], II xi, § 155).[20] This seems relevant for my present purposes. For what makes (TCI) credible is precisely one kind of seeing-as, whereas what makes (TCI) challengeable is another kind of seeing-as. Let me call the first notion *illusory seeing-as*[21] and the second *organizational seeing-as*. On the one hand, as I said at the very beginning, cases of perceptual illusions paradigmatically support (TCI). Such illusions are instances of illusory seeing-as, which is a non-factive sense of seeing: Illusorily seeing something as *F* does not entail seeing that that very something is *F,* hence that that something is *F* (Cf. Mulligan 1988, p. 142). For example, illusorily seeing two segments as different in length does not entail seeing that such segments differ in length, hence that they are such. Yet on the other hand, organizational seeing-as defies (TCI): the experiences of the duck–rabbit figure, or of the famous picture of a dalmatian for that matter, are, as Wittgenstein again says, 'half visual experience, half thought' (2009[4], II, xi § 140), insofar as they involve the lighting up of aspects which may well involve concepts.

However, such seeing-as differences hardly prompt a different stance with respect to (TCI). As I said at the very beginning, one may redescribe the seeing-as experiences involved in the duck–rabbit cases in terms of different seeing-in experiences, in which in virtue of literally seeing a certain figure, one nonliterally sees either a certain subject (a duck) or another subject (a rabbit) in it, respectively. In this respect, one may well claim that the fold of the twofold pictorial experience of seeing-in consisting in nonliterally seeing a certain subject in a picture is precisely a case of seeing-as, namely a seeing the picture as that very subject. Now, this seeing-as is surely a case of illusory seeing-as, for the picture obviously is not identical with the subject it is seen as. Moreover, one may well claim that such a fold is grounded in an experience of organizational seeing-as concerning the picture itself; this is precisely the experience that allows grouping the picture's elements in a certain way.[22]

Second, I have limited myself to considering cases of picture perceptions, in which a two-dimensional entity is seen as pictorially representing a three-dimensional entity. Yet Gestalt switches of this kind also occur when a subject faces a

[19] Along with many others, Nanay (2011, p. 560) holds that attention here may well be involuntary. If this amounts to saying that i) voluntary attention is endogenous attention and involuntary attention is exogenous attention and ii) the endogenous/exogenous distinction is meant as I did before (i.e. not as a distinction between an indirect and a direct way for attention to be mobilized but as a distinction between different forms for helping attention to perform the same grouping job), I utterly agree with him.

[20] See also Walton: 'the problem of the nature of depiction is, at bottom, the problem of the nature of the relevant variety of seeing-as' (1990, p. 300).

[21] This is what Hermerén (1969, pp. 4–8) labels *as-if seeing-as*.

[22] I have defended both claims in Voltolini (2012a).

three-dimensional entity. Consider a three-dimensional instance of the Necker cube. Unlike a case involving a two-dimensional figure, which may be seen either as (a picture of) a cube with a certain face in the foreground and another face in the background or as (a picture of) a cube with those faces in the reverse position, the three-dimensional entity can be directly seen either as a cube with a certain face in the foreground and another face in the background or as a cube with those faces in the reverse position. Now, insofar as concepts may be mobilized in inducing this shift, one may wonder whether ordinary perception of three-dimensional entities also constitutes a challenge to (TCI). What if while wandering around I came across a three-dimensional something that now looks to me as a rabbit? Could not it also look to me as a duck? (Cf. once more Wittgenstein 2009[4], II xi, § 138).

To sum up. Picture perception is cognitively *penetrable*. Concepts *may* induce grouping that produces a phenomenal change in the experience—from the experience of a mere figure to the experience of a *picture,* the depictive representation of something, or from the experience of a figure to the alternate experience of one picture and of another one. This also amounts to a change in its content: the content of a certain 'seeing-in' experience or the contents of two different 'seeing-in' experiences, respectively. As Wittgenstein once said, 'that is why the lighting up of an aspect seems half visual experience, half thought' (2009[4], II, xi § 140). Yet in order for (TCI) to be ruled out, it is enough to show that a conceptual difference prompts a phenomenal difference, independently of whether it also determines an intentional difference.

Acknowledgments A preliminary version of this chapter has been presented to the 2010 Conference of the European Society for Philosophy and Psychology, Universities of Bochum and Essen, August 25–28, 2010. I thank all the participants for their very stimulating questions. Let me also thank Diego Marconi and Alfredo Paternoster for their important comments.

References

Chisholm R (1993) Act, content and the duck-rabbit. In: Canfield JW, Shanker SG (eds.) Wittgenstein's intentions. Garland, New York, pp 94–95

Fodor J (1983) The modularity of the mind. MIT Press, Cambridge

Gombrich E (1960) Art and illusion. Phaidon, London

Hermerèn G (1969) Representation and meaning in the visual arts. Scandinavian University Books, Lund

Hopkins R (1998) Picture, image and experience. Cambridge University Press, Cambridge

Lamme VAF (2003) Why visual attention and awareness are different. Trends Cogn Sci 7:12–18

Levinson J (1998) Wollheim on pictorial representation. J Aesthet Art Crit 56:227–233

Macpherson F (2006) Ambiguous figures and the content of experience. Nous 40:82–117

Macpherson F (2012) Cognitive penetration of colour experience: rethinking the issue in light of an indirect mechanism. Philos Phenomen Res 84:24–62

McIver Lopes D (1996) Understanding Pictures. Oxford University Press, Oxford

Mulligan K (1988) Seeing as and assimilative perception. Brentano Stud 1:129–152

Nanay B (2011) Ambiguous figures, attention, and perceptual content: reply to Jagnow. Phenomenol Cogn Sci 10:557–561

Orlandi N (2011) Ambiguous figures and representationalism. Phenomenol Cogn Sci 10:307–323
Peacocke C (1983) Sense and content. Clarendon Press, Oxford
Peterson MA, Kihlstrom JF, Rose PM, Glisky MA (1992) Mental images can be ambiguous: reconstruals and reference-frame reversals. Mem Cognit 20:107–123
Pylyshyn Z (2003) Seeing and visualizing. MIT Press, Cambridge
Raftopoulos A (2009) Cognition and perception. MIT Press, Cambridge
Raftopoulos A (2011) Ambiguous figures and representationalism. Synthese 181:489–514
Tye M (1995) Ten problems of consciousness. MIT Press, Cambridge
Voltolini A (2012a) How to reconcile seeing-as with seeing-in (with mimetic purposes in mind). In: Currie G, Kot'atko P, Pokorny M (eds) Mimesis: metaphysics, cognition, pragmatics. College Publications, London, pp 383–407
Voltolini A (2012b) Toward a syncretistic theory of depiction. In: Calabi C (ed) Perceptual illusions: philosophical and psychological essays. Palgrave Macmillan, Basingstoke, pp 164–191
Voltolini A (2013) The mark of the mental. Phenomenol Mind 4:124–136
Walton KL (1984) Transparent pictures. Crit Inq 11:246–276
Walton KL (1990) Mimesis as make-believe. Harvard University Press, Cambridge
Wittgenstein L (1980) Remarks on the philosophy of psychology, I-II. Blackwell, Oxford
Wittgenstein L (2009^4) Philosophical investigations. Blackwell, Oxford
Wolfe JM, Oliva C, Horowitz TS, Butcher SJ, Bompas A (2002) Segmentation of objects from backgrounds in visual search tasks. Vision Res 42:2985–3004
Wollheim R (1980^2) Seeing-as, seeing-in, and pictorial representation. In: Art and its objects. Cambridge University Press, Cambridge, pp 205–226

Chapter 16
Singular Thoughts, Seeing Doubles and Delusional Misidentification

Philip Gerrans

Abstract In this chapter, I will suggest (i) that Kevin Mulligan has given a powerful analysis which suggests that the descriptive account of perception is incomplete: We perceive not only properties of objects but objects themselves, (ii) that problems for descriptive theories and the solutions identified by philosophers such as Mulligan (following, among others, Husserl; see Mulligan and Smith, Grazer Philos Stud 28:133–163, 1986; Mulligan, West Ont Ser Philos Sci 62:163–194, 1999) are the basis for contemporary cognitive theories of object tracking, (iii) that theories of object tracking help explain the phenomenology of delusional misidentification syndromes (DMS). DMS are best explained on the assumption that we perceive objects, not just their properties. The objects in question are *selves*. The claim defended here is that when we see a familiar face we see *a particular person,* not merely an assembly of facial features from which we infer the identity of their owner. The way in which we see that person is the same way in which we *see an object in virtue of its perceptual appearance.*

Keywords Delusions · Capgras delusion · Singular reference · Identity · Object tracking · Person files

16.1 Introduction

In this chapter, I extend some ideas about the relationship between perceptual content and demonstrative reference developed by Kevin Mulligan into an unfamiliar area: the explanation by cognitive scientists of delusional misidentification syndromes (DMS). At face value the link is not obvious. In DMS, a patient might say 'my father has been replaced by an imposter' (Capgras delusion); 'I am constantly being followed by a stranger disguised as my father' (Fregoli delusion); 'the person I am looking at is transforming into another person' (delusion of intermetamorphosis) (de Pauw and Szulecka 1988; Spier 1992; Ellis et al. 1994; Ellis 1998; Breen et al. 2000a, b). How can an account of the relationship between perception and reference be relevant to the explanation of these disorders?

P. Gerrans (✉)
University of Adelaide, Adelaide, Australia
e-mail: philip.gerrans@adelaide.edu.au

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_16,

The answer lies in the fact that all these disorders involve a mismatch between the perceptual representation of a face and *the individual to whom that face belongs*. On most theories that mismatch is produced by malfunction in a cognitive system which maps representations of facial features to a representation of the individual to whom they belong. In the Capgras delusion, a familiar face is not matched to the right individual; in the Fregoli delusions, a series of different faces are intractably matched to the same individual and in intermetamorphosis the same face is mapped to a series of individuals. Importantly, that mapping is a quasi-perceptual, cognitively impenetrable, process (Ellis and Young 1990; Stone and Young 1997). The words of the patients report *experiences* of mismatch between appearance and identity which are produced by a face-recognition system whose functioning cannot be altered from the 'top-down' (Bayne and Pacherie 2004). This perceptual aspect of the DMS is the basis for cognitive theorising about their aetiology.

Thus, these disorders raise in acute form a quite general problem faced by theorists of perception. Namely, how we identify *individuals* on the basis of their perceptually presented features (Bedford 2001). The problem is that perceptual experience, on some accounts, presents us not with representations of individual objects but with representations of their properties. For example, seeing a yellow-billiard ball consists in seeing the properties of being spherical, yellow, shaded in a certain pattern. But these are all properties, which are not unique to that particular ball, they are in principle properties which could attach to any object. Thus, it seems that perception presents us not with representations of individual objects (*this* billiard ball) but with representations of the collections of properties (spherical, yellow, partly shaded, etc.). On this *descriptive* account of perception we never directly perceive individual objects, rather we perceive bundles of properties from which we *infer* the existence of an object which instantiates them. This inference may be tacit and cognitively impenetrable, but it is an inference nonetheless.

Using a distinction familiar from metaphysics we can say that if descriptive theories are true perception tracks are *qualitative* not numerical *identity*.

The descriptive view of perception goes naturally with some theories in cognitive science which treat perception as feature detection (Treisman 1998). On these views, perceptual systems keep track of features of the environment, constantly updating a description of the perceptual scene according to the flow of ambient information detected by sensory arrays.[1]

That perceptual systems are feature detectors is not in dispute: A crucial question, however, is whether that is *all* they are. If perception is exhausted by feature detection then we are never in direct perceptual contact with *objects* but only with bundles of co-occurring properties. In the remainder of this chapter, I will suggest (i) that Kevin Mulligan has given a powerful analysis which suggests that the descriptive account of perception is incomplete: We perceive not only properties of

[1] It is not necessary to this descriptive view that representations produced by feature detecting systems are linguistic. Properties might be represented simply by covariation with internal properties of the feature-detecting system. The important point is that the covariation tracks changes in properties not individuals.

objects but objects themselves, (ii) that both the problem for descriptive theories and the solutions identified by philosophers such as Mulligan (following, among others, Husserl; see Mulligan and Smith 1986; Mulligan 1999) are the basis for contemporary cognitive theories of object tracking, (iii) theories of object tracking help explain the phenomenology of DMS. DMS is best explained on the assumption that we perceive objects not just their properties. The objects in question are *selves*.

The claim defended here is that when we see a familiar face we see *a particular person,* not merely an assembly of facial features from which we infer the identity of their owner. The way in which we see that person is the same way in which we *see an object in virtue of its perceptual appearance.*

Precisely, how we do this is the difficult question faced by theorists of perception. The question becomes very important to cognitive neuropsychiatrists trying to understand DMS because the form the answer takes will dictate the conceptualisation of the disorder. In the remainder of this chapter, I suggest that work such as Mulligan's provides the right conceptual framework.

16.2 Preliminary Distinctions

Before we proceed, we need to distinguish three versions of descriptive theories: metaphysical, semantic and perceptual. The metaphysical theory we might call the bundle theory of objects. This is the view that there are no objects which instantiate properties, only collections of properties. On reductive analysis, the yellow-billiard ball turns out to be the co-occurrence of a bundle of microphysical properties and their relations which produce the perceptible properties of yellowness, sphericality, etc. Strictly speaking there is no billiard ball per se just a collection of properties which co-occur more or less reliably (Lowe 1992). The most famous version of this view is of course Hume's theory of personal identity, in which the persisting self is nothing more than a bundle of psychological states. There is no self qua enduring object which exists through time, just a bundle of causally connected psychological states. We never experience *ourselves,* just a flux of psychological properties.

The metaphysical thesis is a view about the ultimate nature of reality and could be true or false independently of the semantic and perceptual versions of descriptive theories of objecthood. It seems to be the case that a fundamental feature of language is *reference* to individual objects. When we talk or write about a billiard ball or a tiger changing colour or shape or a person growing older we refer to an individual object which remains constant while its properties change. Theories of reference cannot analyse this phenomenon away and it persists as an ineliminable feature of language, irrespective of how metaphysical controversies about the nature of objects are resolved.

Similarly, whether perception involves the representation of individual objects is a question which cannot be resolved independently of both the metaphysical and semantic questions. It could be the case that there are no objects, and it might be the case (at least for some objects with a uniquely identifying set of properties

in the actual world) that descriptive theories of reference are sufficient to explain singular reference. But the metaphysical and semantic adequacy of descriptive theories would not automatically resolve the question of whether we *see* objects or properties. Humans might be engineered to get around their world by perceiving it as populated by clusters of nonaccidental regularly co-occurring properties, *or* by perceiving objects which instantiate those properties. Evolution might be blind to metaphysical and semantic distinctions.

Metaphysics and semantics, however, provide hypotheses about the nature of objects and properties which can frame inquiry into the psychology of perception. For example, suppose that we do in fact succeed in parsing the structure of perception sufficiently to determine whether in fact objects or properties are represented in perceptual experience and that we conclude that we perceive objects. What are these objects and how can we see them given that they are always detected in virtue of their perceptible properties? Here, metaphysics offers a menu of possibilities. Objects might be represented in experience as haecceities (the property of being this particular thing), bare particulars, individual substances, substance sortals or as components of irreducible object property complexes.

16.3 The Inadequacy of Descriptive Accounts

Mulligan proposes a solution to the problem of perception as part of a solution to the problem of singular reference. In fact, given the dependence of singular reference (that is the process of naming or indicating a unique individual) on perception it is not surprising that the structure of the problem for descriptive theories of reference and descriptive theories of perception appears the same:

> It would a priori be very surprising if an account of the way language works were to be independent of an account of perception. Perception and language are two of our most basic capacities. Two features of linguistic behaviour, at least, are common to humans and many other animals: expression or indication and signalling or steering (the dances of bees) and these are inseparable from perception. In very many ways, representation, the coordination of words and objects, grows out of and relies on, indication and steering. (Mulligan 1997)

The point is that the ability to refer to individuals is not something which is created by language. Rather singular reference depends on the prior perceptual ability to pick out individuals. In fact Mulligan argues, *only* those linguistic expressions whose use essentially involves perception can succeed in singular reference: that is to a unique individual.

We can see this when we look at the way descriptive terms refer. It is commonplace that descriptive reference which predicates properties of individuals cannot succeed in picking out unique individuals. ‘The man standing in the corner wearing glasses’ is a description potentially satisfied by more than one individual. Making the description more precise does not solve the problem since it only adds another descriptive component.

Thus, the difficulty for descriptive theories is, in effect, that definite descriptions can only ever provide conditions for establishing qualitative identity. That is to say

that two individuals whose properties are identical can satisfy the description. But when we use the name 'Castor' we want that name to refer to Castor not his twin, Bollux. If, however, we substitute a definite description for 'Castor' then we also refer to Bollux.

The first step in the solution is to appeal to something like a demonstrative sense: the linguistic equivalent of pointing to *this particular man*. But of course that alone cannot complete the task of securing reference. As Mulligan points out, the only way to provide the link is to complete that demonstrative by linking it to an episode of perception. 'The one we are looking at now':

> Let us suppose, to begin with, that demonstratives, like proper names, have a sense which is simple and is grasped or instantiated by the speaker. This sense is incomplete. What completes it? *Veridical perceptual content*. (my italics) (Mulligan 1997)

If I point to the corner and verbally direct your attention to a man wearing glasses then I succeed in referring to him, provided you correctly see him and attach my description to the man you see. The description is supplemented by the demonstrative indication *which is in turn completed by the perceptual identification of the man in question*.

But as Mulligan then points out, in order for perception to anchor singular reference in this way the content of the perceptual episode must pick out an individual. Otherwise the explanation is circular: If perceptual content is descriptive then it cannot solve problems for descriptive theories of reference.

Mulligan draws the distinction between descriptive and nondescriptive in terms of a distinction between conceptual and nonconceptual. By conceptual he means able to be the object of a judgement (or a component of a thought which can be the object of judgement). Conceptual content is detected by its intensional logical properties: Sameness of conceptual content of two referring expressions does not guarantee the numerical identity of the objects referred to. Two expressions with the same intension do not necessarily have the same extension. Similarly, two perceptual episodes with the same intension or conceptual content ('yellow', 'spherical') do not necessarily have the same extension (this particular billiard ball). Thus, conceptual content is descriptive in the sense we are using it.

Consequently, there must be some aspect of perceptual content which is nonconceptual (or as he might put it purely extensional) if an episode of perception is to be genuinely singular. We must be able to see *this thing* not merely those of its properties, which can also be borne by another identical or similar thing. As Mulligan puts it, the content of the perceptual demonstrative must be nonconceptual.

Mulligan quotes Husserl on this point:

> I say 'this', and now mean the paper lying before me. Perception is responsible for the relation of my word to this object, but my meaning does not lie in perception. An act of this-meaning builds itself on my perception, depends on it. Without the perception—or some correspondingly functioning act—the pointing would be empty, without definite differentiation. For the indeterminate thought of the speaker as pointing to something…is not the thought we enact in the actual pointing. (Husserl LI, VI, § 5) (Mulligan 1997, p. 126)

Thus, if perception is to anchor singular reference the content of perception cannot be descriptive, on pain of circularity.

16.4 Seeing Things. Object Tracking

A crucial point made by Mulligan is that when we turn our attention from language to perception we turn from a phenomenon (semantic theory) often conceptualised as static and context independent to one which is dynamic. For semantic theory, the problem is dramatised by indexicals and demonstratives. A context-free rule of reference for such singular expressions is always incomplete. If 'here' refers to a particular place then a rule such as '"here" refers to the place where it is uttered' cannot secure reference. It needs to be supplemented on the occasion of use by some way of tracking the actual location: 'This place'. But the demonstrative is also incomplete unless we can somehow link it to perception of the location.

Which means, in fact, that perception must function in a dynamic way: It must be able to track objects independently of any descriptive content in order to serve as an anchor for descriptive content. This must be the case because descriptive content is constantly changing: As we move around the room while fixing our gaze on an object the descriptive content of perception changes.

Perceptual theorists thus note that one problem faced by perceptual processing is to construct what they call a 'structural description' of the object which preserves the context invariant properties. Thus, we continue to see constancies in colour and shape despite fluctuations in the retinal information.

However, as we saw above, identity even of such abstract descriptive content, does not necessarily secure identity of objects perceived. If we switch our gaze among apparently identical objects we need to be able to detect whether we are seeing one or many.

As perceptual theorists put it:

> we are constantly confronted by informational samples which originate from different times and places, both within and across modalities. In these cases we need to determine whether these samples comes from the *same object undergoing a change* (e.g. of location, colour or some other perceptible property) *or from different objects*. (Bedford 2001)

Perceptual scientists have investigated this phenomenon in depth, confronting subjects with arrays of moving objects and varying their properties such as colour, size, shape location and trajectory while occluding and unmasking them during a short scenario (Pylyshyn 1984, 2001; Haladjian and Pylyshyn 2006). Subjects are asked to keep track of the objects during the scenario, a task which implies an implicit representation of the distinction between numerical and qualitative identity and other experimenters have reached to two important conclusions. The first is that spatiotemporal continuity is insufficient for the representation of numerical identity. Instead, the perception of a set of properties at a location causes the representation of an object which then is tracked through its changes in location and appearance. As long as the objects stay 'bounded' people tend to judge that they are the same object. The second conclusion is that the *identity of objects is represented in perceptual experience.*

Pylyshyn describes the route to this conclusion, in terms, strikingly reminiscent of Husserl and Mulligan:

> If perceptual representations are to be grounded in the physical world then a causal link is essential at some stage in the process. The usual link that has been assumed is a semantic one—the objects that fit a particular description are the ones picked out and referred to. While this may be generally true this cannot be the whole story since it would be circular. The symbolic description must bottom out—must be grounded in objects or properties in the perceptual world. Recent evidence has suggested that the grounding is done in *objects rather than properties*. (Pylyshyn 2007) (My italics)

When we attend to a set of features our perceptual experience includes a representation of the particular object which has those features. Perception, scientists call the representation of the object which remains in existence while properties change an 'object file'. Metaphysicians call such a representation a 'substance sortal' which can be defined as the representation you need in order to count identical objects such as skittles, billiard balls or clones. To do so you must be able to represent the distinction between qualitative and numerical identity (Wiggins 1997; Wiggins 2001).

Normally, perceptual appearances are mapped to object files which anchor further perceptual and higher-level processing. Consequently, when we attend to objects we experience them as *particular things* (objects) which *appear a certain way* (i.e. have properties). There is a debate within vision science about whether object files are created in very early visual processing, in order to coordinate automatic sensorimotor actions such as grasping, or whether they emerge when visual attention is driven from the top down to provide information for higher-level integration. Pylyshyn argues for the former view but for our purpose this debate does not need to be resolved. The important point for this discussion is that on either conception *the identity of objects is experienced*. If this is the correct explanation of the phenomenology then appearance and identity are dissociable elements of perceptual experience.

Note also that if Pylyshyn and others are correct then there is empirical support for the idea that perception can anchor singular reference via nondescriptive (nonconceptual) content which identifies objects.

16.5 Seeing People: Delusional Misidentification

I noted at the outset that a consideration in favour of the idea that we represent objects is ecological. We need to represent the world as populated with objects in order to engage successfully with it. Perhaps, the most important objects for humans are other people. It matters crucially that we correctly identify other people and our ability to do is quite amazing. People are easily and effortlessly identified under very adverse conditions, in poor light, at a distance, after years of aging or cosmetic surgery.

Theories of face recognition also face the problem for descriptive theories. The face-recognition system confronts a face whose properties (eye colour, shape of nose, relationship between features, etc.) are the basis of identification. It must map

these properties to a particular person in order to identify the owner of the face (Young and Burton 1999; Schweinberger and Burton 2003). Note that it is not sufficient to map these perceptual properties to a name, since the name in that context becomes merely another property to add to the description ('is called Bollux'). The problem is that no such description is sufficient to identify a unique individual who bears all and only those properties. It is always logically possible at least that some other individual also bears those properties.

I suggested above that if the object tracking theory of Pylyshyn is correct then appearance and identity can potentially dissociate in perceptual experience.

Delusions of misidentification seem to be instances where this possibility is realised in the experience of recognising faces. In these cases, subjects see a familiar person but say that they see someone else. Or they see an unfamiliar person but say that they are a familiar person. What seems common to these delusions is an experience in which identity and appearance dissociate.

In what follows, I discuss the Capgras delusion but the account I develop is equally applicable to other DMS. In fact, one advantage of this account is that it seems that it is the only one with the potential to unify the different DMS.

Although the literature is vast, I concentrate on a recent exemplary study by Brighetti et al. who studied a patient, *YY,* with Capgras delusion (Brighetti et al. 2007). The patient was in some ways atypical because she showed no anatomical deficits or lesions. However, her delusion was entirely typical. It followed an incident in the classroom when she was unable to read her own writing (which suggests a lack of recognitional ability for familiar stimuli). This was followed by an episode of catatonia which led to hospitalisation. Following that episode, she showed reduced emotional warmth to her family and then called the police claiming her father had been replaced by an imposter. For the next 2 months while under psychiatric treatment, she failed to acknowledge six family members and her professor although she had no difficulty identifying other familiars. Eventually, the delusion subsided for all except her father, who she continued to try and unmask as an imposter.

Brighetti et al. tested this patient on photographs of family members (including the father), familiar and unfamiliar faces and neural objects.

Her eye movements were monitored for frequency, location and duration of fixations while she looked at the photos. An interesting and possibly distorting aspect of tests like this is that photographs present static faces, which must affect the scanning process since inferring information from facial expression and gaze direction involves circuits such as the superior temporal sulcus which respond to *expressive movements*.

Her skin conductance response (SCR) was also monitored. SCR is an index of amygdala activation which sets up autonomic response senses as affective feelings. This is important because a standard account of this delusion explains it in terms of loss of affective response to familiar faces. The basic idea of standard accounts is that seeing a familiar face produces an affective response which is used to identify the familiar person (Bauer 1984; Ellis 1986; Tzavaras et al. 1986; Tranel and Damasio 1988; Bruyer 1991; Young and de Haan 1992; Ellis et al. 1993; Young and

Burton 1999; Breen et al. 2000a; Breen et al. 2001; Ellis and Lewis 2001; Lewis and Ellis 2001; Schweinberger and Burton 2003). When that affective response is absent due to damage to circuits which link the amygdala to the face-recognition system, centred on the right fusiform gyrus, we see a familiar person but we do not experience the normal affective response. The delusion is an attempt to explain that experience.

One feature of this standard account is that it is *essentially a descriptive account of person identification*. It assumes that people are identified by facial appearance (a set of perceptual properties) together with an affective response. The affective response is in effect the final property which completes the identifying of definite description.

However, this standard account also predicts that conditions in which amygdala functioning or connectivity is impaired should result in misidentification experiences and this prediction does not seem to be born out.

YY, in fact, showed almost no difference in SCR (taken to be an indicator of amygdala activation which produces an affective response) for familiar and unfamiliar faces, unlike controls who showed the normal increase in SCR for familiars. *YY* was able to identify all familiars including those for whom she had previously had the delusion and the father for whom the delusion was maintained. This is consistent with the idea that differences in amygdala activation are not the essential element in identification.

However, *YY's* scanning of faces differed markedly from that of controls. Controls fully explored the faces of both familiar and unfamiliar people. *YY,* however, showed reduced exploration relative to controls for both familiar and unfamiliar faces. Furthermore, *YY* showed reduced exploration of familiar faces relative to unfamiliar. Not only that but she showed a different pattern of exploration of faces for those faces which had been the subject of the delusion. Specifically, she did not avoid the eyes.

Interestingly, the patient correctly identified all familiar faces from the photographs even though she retained the delusion for her father.

Brighetti et al. drew some conclusions. First, familiarity, accompanied by SCR, normally leads to more elaborate exploration in the eye region. This is consistent with the idea that exploration of the eye region is essential for inferring social information such as emotion and intention.

However, some patients who have reduced amygdala activation avoid exploration of the eye region. These patients produce the Capgras delusion. As they put it: 'identity recognition of familiar faces associated with a lack of SCR results in gaze avoidance of the eye region.' (196). Brighetti et al. are suggesting that the delusions results from conflicting information. 'This face belongs to *X*' and 'this face is unfamiliar'. In effect, the subject identifies the seen face but does not recognise her as familiar.

This suggests that amygdala activation is not required to establish the exact identity of the face. Otherwise, the *inconsistency* between identity and unfamiliarity would not be invoked to explain avoidance. Furthermore, if amygdala activation is necessary to establish the exact identity of the face *YY* and the many other patients who lack SCR would misidentify faces.

Initial versions of the delusion suggested that identity was computed on the basis of familiarity (indexed by amygdala activation which produces affective response to familiars) and appearance. When familiarity is absent, identity cannot be computed and the Capgras delusion results. It is, however, difficult to extend this account to the other delusions of misidentification. For example, if the Fregoli delusion is a consequence of the sense of presence being mismatched to the appearance of a stranger the Fregoli delusion should take the form of a belief that the subject is being followed by a stranger in disguise, not a specific person. Similarly, in delusions of intermetamorphosis if the cause is the waxing and waning of a nonspecific sense of presence while the appearance of the target remains constant the delusion should not report the experience of seeing the target transform into a series of different selves.

Early versions of the standard account recognised this problem and tried to solve it by suggesting that the affective response was specific to particular identities.

Brighetti's account, together with the other evidence we have considered, suggests that identity is matched to appearance *independently of amygdala activity and prior to any abductive process*. This matching process produces experience of identification or misidentification. But how?

There must be such a mechanism because in normal cognition all the properties represented by different elements of the system: appearance, familiarity, semantic information, name are attributed to *the same person*. This representation of identity *anchors* the attribution of different properties.

The solution is to import the concept of an object file into the architecture of face recognition. And in fact, as we noted above, the same considerations which argue in favour of object files also apply to persons. They are things which persist through time despite changes in their properties and can be identified, reidentified and counted. The concept of a person as an enduring entity which undergoes physical and mental changes is essential to human cognition precisely because so much of it revolves around the tracking of identity. Consequently, we need a way to represent a person as an entity independently of her appearance. As Erana et al. put it: 'keeping track of agents seems to require some sort of mechanism for the selection of individuals, the creation of a referential link and its maintenance over time' http://www.interdisciplines.org/objects/papers/3.

The question then arises whether the ability to represent persons as numerically distinct individuals which bear properties is a high-level conceptual ability or something much closer to a perceptual ability. If it is the latter, then as with ordinary objects we should be able to represent the identity of distinct persons in experience. In effect, when we see a person we would map appearance to a 'person file' which stands in the same relationship to appearances as object files to appearances of everyday objects.

Erana et al. have proposed that such 'person files' are represented by automatic processes rather than high-level controlled processes. They call them 'agent files'. Representations, which help keep track of perceptible features of conspecifics such as animacy, expressive bodily movement manifesting intentions and emotional expression. These features of the world are processed by specialized systems which

develop early in life. Erana et al. suggest that the integration of outputs of these systems depends on agent files. The infant does not see the expression of concern, a reaching motion, a body. She sees her mother, *this person,* reaching towards her. This hypothesis is 'an extension of the studies on object individuation and tracking to the domain of perceptual individuation and tracking of entities endowed with agency'.

Erana et al. argued that the concept of an agent file or person file provides a parsimonious explanation of many aspects of infant social cognition, in just the same way as the concept of an object file integrates findings about infant perception and numerical cognition with studies of adult object tracking.

The concept of a person file also economically explains DMS and a range of other conditions which involve misidentification. In particular, the concept of person files explains a subtle distinction between two ways of misidentifying objects and persons which is important to delusions of misidentification.

The distinction is between qualitatively identical duplicates or replicants and distinct individuals who appear the same. Clones or identical twins are *duplicates*. They have identical intrinsic or essential properties and are qualitatively identical. If the Regius Professor of Gender Studies has gender reassignment surgery and extensive plastic surgery in order to make himself indistinguishable from Hilary Clinton he is not a duplicate of the Secretary of State. The Regius Professor does not share the same intrinsic properties although she is now qualitatively identical to Hilary Clinton. If Hilary Clinton had an identical twin she would be a duplicate or clone.

The distinction is important in the case of ordinary physical objects and places since it explains reduplicative paramnesia, in which people say that objects have been duplicated or multiplied rather than replaced or substituted.

Interestingly, some early neuropsychological explanations of the Capgras delusion assimilated it to reduplicative paramnesia treating the misidentification experience as an experience of reduplication. For example, Alexander et al. suggested that 'the Capgras syndrome may be a form of reduplicative paramnesia with the same pathologic substrate'. Their suggestion was based on a case of a man who on returning home 10 months after being hospitalised after a severe head injury claimed that his family had been replaced by a new family virtually identical to his own. The only difference was that the children looked a year older (Alexander et al. 1979).

There are in fact some reports of the experience of reduplication for persons. Indeed, some theorist have proposed a separate category of 'Clonal pluralisation of the self' to explain a patient who claims to have four psychologically and physically identical doubles (Vörös et al. 2000). This patient was, however, schizophrenic so the aetiology of the delusion might be very different from the delusions under discussion caused by lesions to the face-recognition system.

However, close attention to the phenomenology of Capgras delusion suggests that it is not a reduplicative phenomenon. Rather, the experience seems to be of replacement by a different person, an imposter or double rather than a duplicate. If Capgras was a version of paramnesia for people we should expect reports of clones or twins.

While the phenomenology is not transparent from the clinical reports it does seem that the distinction between paramnesia/duplication and substitution/replacement is both conceptually coherent and reflected in phenomenology.

The idea that appearances are mapped to person files gives an elegant explanation of the distinction between replication and substitution which can explain the three classic delusions of misidentification. It also provides an explanation of subtle distinctions in the phenomenology of different cases. Rather than three factors: appearance, familiarity, semantic information being the basis for identification there are now four. Numerical identity is provided by the person file which organizes the integration of the other three types of information.

If this is correct then there may be cases in which representation and integration of any of these elements due to malfunction in the relevant circuitry. These malfunctions can be transient as in fleeting *deja* or *jamais vu* experiences but in the DMS they are produced by more persistent failures of the system.

On this account, we would predict the fractionation of the phenomenology of identification along different dimensions according to the way different elements are combined. We could expect cases of hyperfamiliarity or hypofamiliarity for both identified and misidentified faces. And identification could take the form of establishing qualitative identity (all information intact but the person file absent) or numerical identity.

This account integrates the explanation of the classic DMS in terms of a mismatch between appearance and person file associated with absence or presence of familiarity. The Fregoli delusion is a case of the wrong person file or no person file being mapped to a strange face. In intermetamorphosis a series of different person files are activated by the same face.

Reduplicative phenomena can be explained in terms of inappropriate creation of a new person file for a familiar face.

Similarly, the eerie loss of sense of presence or its inverse 'hyperfamiliarity' can be explained by the hypo or hyperactivation of the amygdala by a seen face. Given that the amygdala is activated by early recognitional processes this type of malfunction could arise with or without the activation of the person file depending on the circuitry involved.

Turning to the Capgras delusion, whose explanation is a focus of cognitive neuropsychiatry, there is more than one way it might arise. It seems unlikely that it is a reduplicative phenomenon but it clearly involves a mismatch between appearance and familiarity. On the account above, this could be associated with or without mapping to the correct person file.

Brighetti et al.'s account suggests that the delusion arises as a result of the inconsistency between identification and absence of familiarity which suggests that the correct person file is intact. A case of numerical identity minus familiarity.

Another possibility is that it is a case of the loss of the correct person file, leading to experience of qualitative identity minus familiarity. The consistent reports of imposterhood, violence and suspicion towards the imposter seem to me to support this interpretation. The Capgras subject actually *sees the wrong person,* not a duplicate or the right person minus familiarity.

Given the clinical rarity of DMS and the fact that they often occur in situations which make cognitive theorizing a low priority (Alexander's patient was severely disabled and right hemisphere lesions are often associated with other serious damage) verifying any hypothesis is difficult.

However, there does seem to be one conclusion that we can draw when we combine case studies of DMS with the philosopher's arguments against descriptive theories of perceptual content.

We see people.

References

Alexander MP, Strauss DT et al (1979) Capgras syndrome: a reduplicative phenomenon. Neurology 39:334–339

Bauer RM (1984) Autonomic recognition of names and faces in prosopagnosia: a neuropsychological application of the guilty knowledge test. Neuropsychologia 22:457–469

Bayne T, Pacherie E (2004) Bottom-up or top-down? Campbell's rationalist account of delusions. Philos Psychiatry Psychol 11:1–12

Bedford F (2001) Towards a general law of numerical/object identity. Curr Psychol Cogn 20(3/4):113–176

Breen N, Caine D et al (2000a) Towards an understanding of delusions of misidentification. Mind Lang 15:74–110

Breen N, Caine D et al (2000b) Models of face recognition and delusional misidentification: a critical review. Cogn Neuropsychol 17:55–71

Breen N, Coltheart M et al (2001) A two-way window on face recognition. Trends Cogn Sci 5:234–235

Brighetti G, Bonifacci P et al (2007) Far from the heart far from the eye: evidence from the Capgras delusion. Cogn Neuropsychiatry 12(3):189–197

Bruyer R (1991) Covert face recognition in prosopagnosia: a review. Brain Cogn 15:223–235

de Pauw KW, Szulecka TK (1988) Dangerous delusions: violence and the misidentification of syndromes. Br J Psychiatry 152:91–97

Ellis HD (1986) Processes underlying face recognition. In: Bruyer R (ed) The neuropsychology of face perception and facial expression. Lawrence Erlbaum Associates Inc, Hilsdale

Ellis HD (1998) Cognitive neuropsychiatry and delusional misidentification syndromes: an exemplary vindication of a new discipline. Cogn Neuropsychiatry 3:81–90

Ellis HD, Lewis MB (2001) Capgras delusion: a window on face recognition. Trends Cogn Sci 5(4):149–156

Ellis HD, Young AW (1990) Accounting for delusional misidentification. Br J Psychiatry 157: 239–248

Ellis HD, Young AW et al (1993) Covert face recognition without prosopagnosia. Behav Neurol 6:27–32

Ellis HD, Whitley J et al (1994) Delusional misidentifications: the three original papers on the Capgras, Frégoli and Intermetamorphosis delusions. Hist Psychiatry 5:117–146

Haladjian HH, Pylyshyn ZW (2006) Implicit multiple object tracking without an explicit tracking task. J Vis 6(6):773

Lewis MB, Ellis HD (2001) A two-way window on face recognition: reply to Breen et al. Trends Cogn Sci 5:235

Lowe E (1992) Experience and its objects. In: Crane T (ed) The contents of experience, pp 79–104

Mulligan K (1997) How perception fixes reference. Sprache und Denken Language and Thought, herausgegeben von, edited by Alex Burri Walter de Gruyter, Berlin, New York, 122–138

Mulligan K (1999) Perception, particulars and predicates. West Ont Ser Philos Sci 62:163–194
Mulligan K, Smith B (1986) A Husserlian theory of indexicality 1. Grazer Philos Stud 28:133–163
Pylyshyn ZW (1984) Computation and cognition: towards a foundation for cognitive science. The MIT Press, Cambridge
Pylyshyn ZW (2001) Visual indexes, preconceptual objects, and situated vision. Cognition 80 (1–2):127–158
Pylyshyn ZW (2007) Things and places: how the mind connects with the world. The MIT Press, Cambridge
Schweinberger S, Burton M (2003) Covert recognition and the neural system for face processing. Cortex 39:9–30
Spier S (1992) Capgras syndrome and the delusions of misidentification. Psychiatric Ann 22: 279–285
Stone M, Young AW (1997) Delusions and brain injury: the philosophy and psychology of belief. Mind Lang 13:327–364
Tranel D, Damasio AR (1988) Non-conscious face recognition in patients with face agnosia. Behav Brain Res 30:235–249
Treisman A (1998) Feature binding, attention and object perception. Philos Trans R Soc Lond B Biol Sci 353(1373):1295
Tzavaras A, Luauté JP et al (1986) Face recognition dysfunction and delusional misidentification syndromes (DMS). In: Ellis HD, Jeeves MA, Newcombe F, Young, AW (eds) Aspects of face processing. Martinus Nijhoff, Dordrecht
Vörös V, Tènyi T et al (2000) Clonal pluralization of the self: a new form of delusional misidentification syndrome. Psychopathology 36(1):46–48
Wiggins D (1997) Sortal concepts: a reply to Xu. Mind Lang 12(3–4):413–421
Wiggins D (2001) Sameness and substance renewed. Cambridge Univ Press, Cambridge
Young AW, Burton AM (1999) Simulating face recognition: implications for modelling cognition. Cogn Neuropsychol 16:1–48
Young AW, de Haan EHF (1992) Face recognition and awareness after brain injury. In: Milner AD, Rugg MD (eds) The neuropsychology of consciousness. Academic Press, London, pp 69–90

Chapter 17
Reconstructing (Phenomenal) Consciousness

Alfredo Paternoster

Abstract In this chapter, I shall discuss Block's distinction between phenomenal consciousness and access consciousness. I will argue that although Block's proposal has the merit of accounting for some important distinctive phenomena, it should nonetheless be given up, in favor of a single, graded notion of consciousness. There is only one consciousness, which one can possess in different degrees.

Keywords Phenomenal consciousness · Access consciousness · Qualia · Blindsight

17.1 A Quasi-Standard Distinction

It is well known that the words "conscious" and "consciousness" have several uses. For instance, "conscious" can be predicated both of a person ("she is severely injured, but is conscious") and of a mental state ("she has a conscious experience of redness"); and the property of being conscious can be directed to something ("she is conscious of seeing a red blob") or not; the latter distinction is also expressed by saying that consciousness can be transitive or intransitive. Indeed, the concept expressed by the word "conscious" is, to say the least, a "cluster concept": Different cases fall under the same concept but these differences are subtle and elusive. For this reason, many authors have tried to clarify the concept by drawing some distinctions. Ned Block's distinction between *phenomenal consciousness* (P-consciousness) and *access consciousness* (A-consciousness) has probably been the most influential. This chapter is mainly devoted to a discussion of Block's distinction. I will argue that although Block's proposal has the merit of accounting for some important distinctive phenomena, it should nonetheless be given up, in favor of a single, graded notion of consciousness.

Let us start, then, by quoting Block's distinction (1994, p. 214; see also 1995, p. 231):

- A mental state is A-conscious if and only if its content: (a) is freely available as a premise in reasoning; (b) is poised for rational control of action, and (c) is poised for rational control of speech. Note that although the ability to report the content

A. Paternoster (✉)
University of Bergamo, Via Pignolo 123, 24121 Bergamo, Italy
e-mail: alfredo.paternoster@unibg.it

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_17,

of one's own state is certainly not the most important among these features, it is nonetheless the best empirical criterion to establish whether a mental state is A-conscious. Self-consciousness could be regarded as a particular case of A-consciousness (even if, in Block 1994, self-consciousness and A-consciousness are distinguished—this is not at all relevant to my discussion).

- A mental state is P-conscious if and only if it is a state experienced in the first person. That is to say, "P-consciousness is just experience" (Block 1994). As many, following Nagel (1974), put it, there is something it is like to be when someone is in a P-conscious state.

Block's distinction has played an important role in the debates about the possibility of giving a scientific explanation of consciousness, and was partly motivated by the assumption that P-consciousness escapes functional treatment. Indeed, if one takes seriously Block's distinction between P-consciousness and A-consciousness, then he will be inclined to endorse the view that P-consciousness cannot be scientifically accounted for. Or, at least, there is a very large consensus on the thesis that it is P-consciousness that raises the hard problem in explaining consciousness. Thus, one might even argue that Block's distinction is part of the problem, insofar as it offers some prima facie reasons to think that there is a certain kind of consciousness that cannot be scientifically studied—there is no way to address subjectivity in a scientific way. I will come back to this point in the next section.

However, despite the fact that Block's distinction has been quite popular and is admittedly supported by both conceptual and empirical considerations, I think we should resist this idea of a "dual consciousness" and try to sketch instead a *unique* notion of consciousness. Indeed, I will argue that Block's distinction, on the one hand, can be defeated from an ontological point of view, and, on the other hand, does not give us any explanatory payoff. I will put forward a single concept of consciousness wherein the phenomenal aspect is somewhat prior—there is no consciousness without phenomenal effects—but, at the same time, involving some of the aspects that Block subsumed under the head of A-consciousness. In other words, my position can be sketchily described by saying that *consciousness is fundamentally phenomenal but there is no phenomenal effect at all without some kind of access*.

Consciousness involves first and foremost the availability of "something" to an organism *as a whole* and this implies both a phenomenal effect—a feeling—and, as I shall explain later, some kind of access.

17.2 An Argument Against the Distinction

Let me start by pointing out that Block's distinction is unsatisfactory in a twofold sense.

First, Block predicates the property of being conscious, in both its aspects (*P* and *A*), of mental states, not of persons or subjects. This is, however, quite counterintuitive: I would say that it is *subjects* that are in the first place conscious rather

than unconscious. One can certainly talk about consciousness as a property of a *mental state,* but this use seems to be parasitic on the notion of a conscious subject.[1] This is apparent in the Blockean definition of P-consciousness: A mental state is P-conscious if and only if *there is something it is like to be* when the subject is in that mental state. Also, note that this definition does not provide any insight into the concept of being P-conscious, which is brought back to the pre-theoretical and quite general concept of a conscious subject—simply conscious—without any further specification. Arguably, it is not by chance that when Block proposes a definition of a conscious state which is not parasitic of the notion of a conscious subject, as he does in the case of the definition of A-conscious, the alleged states subsumed under the definition can hardly be regarded as genuine conscious states (see below). Without a reference to the subject, the notion of consciousness seems to vanish.

Second, it is the distinction itself, however reasonable it may be, that gives the impression that P-consciousness cannot be scientifically explained to the extent that the content of a P-conscious state is identified with a collection of "third type *qualia,*" that is, with some alleged qualitative, intrinsically private properties that can be neither functionalized nor communicated. In other words, the widespread belief that there is an irreducible (phenomenal) residue could be exactly produced by the dissociation between access properties and phenomenal properties: If one defines P-consciousness "by subtraction" (of functional properties), then the result he should expect is just something elusive. Postulating an *independent* P-consciousness from the start is starting on the wrong foot or, at least, putting the cart before the horse. As we shall see, if we give up the distinction, then the perspectives for a scientific explanation of consciousness are better.

Of course these considerations are not conclusive per se—they do not amount to an argument. Here is mine.

The basic reason underlying Block's distinction is that our intuitions concerning the appropriateness of taking certain phenomena as conscious rather than unconscious are wavering. And the only way to account for that would be to distinguish two kinds of consciousness. The relevant phenomena would, in fact, amount to cases of dissociations between the two aforementioned kinds of consciousness: A-consciousness without P-consciousness and the other way around. What I am going to do is to discuss these alleged dissociation cases and assess whether we can find a different interpretation for them.

[1] Recently, Simone Gozzano (2009, p. 10) has argued for the opposite view: "Conscious" denotes first a property of a mental state. However, his main reason for this thesis is that saying of someone (a person, or an animal, or, maybe, a robot) that he is conscious does not make any clear sense, and that this way of speaking is actually elliptical—there is a specific, though unexpressed, mental state that we have in mind when we attribute consciousness to someone. I agree on the point that to say that someone is conscious is quite vague; but this does not undermine the fact that what we mean by saying that a mental state is conscious is that the subject is conscious (of something) when he is in that mental state. The vagueness of "conscious" is one thing, the core sense of "conscious" is another. It seems to me that what Gozzano's argument shows is that when we are speaking theoretically (when we are theorizing about consciousness), we do better to talk about conscious (or unconscious) mental states.

Let us consider, first of all, the case of P-consciousness without A-consciousness. According to Block, this can easily be found both in experimental conditions and in ordinary, daily situations. As an example of ordinary situation, take the case of a person concentrating on a certain task who does not realize that there is a noise—is not aware of the noise—but a few minutes or so later, when she eventually comes to be aware of the noise, realizes that the noise was already there before: She heard it before also, but she did not notice it**.** As Block put it, "you were aware of the noise all along, but only at midnight were you consciously aware of it. That is, you were phenomenally conscious of the noise all along, but only at midnight did you become access-conscious of it" (1994, p. 215). There are many other examples of this kind (Tye 2005, p. 2–4).

As an example of experimental case, consider Sperling's (1960) classic experiment. An array of 12 letters (4 × 3) was shown to subjects for 50 msecs; then subjects were required to report which letters they had seen. They were able to tell correctly, in average, four, though they reported having seen many more. However, if, immediately after being shown the array, subjects were alerted to focus on a certain row (for instance, by means of a distinct kind of sound: low = first row: medium = second row, etc.), they were able to report correctly all the letters in the relevant row (for a discussion, see Block 2007).

Block interprets both the experimental situation and the ordinary situation as cases in which persons do not *cognitively* access, either because of lack of attention or because of spatial and temporal limits of the working memory, some information that, however, is phenomenally available.

A problem with this description is that it is not clear what it means that someone should be *phenomenally* conscious of a content that, at the same time, is not "present" to the agent's consciousness, at least in an intuitive sense of consciousness. Admittedly, Block's distinction is conceived of exactly to account for this ambiguity: Something is at the same time present and not present to consciousness. However, as Crane (2002) has argued, these cases can be accounted for more easily, without duplicating the concept of consciousness. For example, they can be described as cases of consciousness without attention. If we accepted, as indeed we shall do, this interpretation, the concept of A-consciousness could turn out to be an idle wheel. Before drawing this conclusion, however, we must also analyze the converse case: A-consciousness without P-consciousness.

And here is the major problem: There are no indisputable cases of this kind. According to Block, the only case of A-conscious and not P-conscious state is the "super-blindsight," a fictitious syndrome very similar to the real blindsight. In real blindsight (e.g., Weiskrantz et al. 1974) subjects affected by (even massive) damage in the primary visual area (V1) are successful in visual discrimination tasks despite being unaware of the presented stimuli: They report being totally blind. They say to the experimenter, with understandable irritation, that they are unable to accomplish the requested task, but, when asked to try anyway, they present surprisingly good performances.

This description seems to fit what Block is looking for: There is a visual content, processed somewhere in the brain, which is available for other cognitive processes, but this content is not subjectively (i.e., phenomenally) conscious. There is some

information that the subject is not subjectively aware of, but that is poised for judgments. However, strictly speaking—so Block points out—the patient has no A-consciousness of the stimulus either, because, until she hears her own guess, she cannot use the information *freely* in reasoning or in rational control of action (Block 1994, p. 215). In fact, Block's definition of an A-conscious state requires that information be *freely* brought to bear on cognitive processes (as we saw above). That is why Block puts forward the case of super-blindsight, wherein the relevant constraint turns out to be satisfied.

Super-blindsight, however, is not real. Subjects need a cue, in order to make actually poised the visual content for their cognitive processes. Thus it seems, after all, that there are actually no cases of A-consciousness without P-consciousness, even if, of course, this situation is conceptually possible.

One might argue, on the other hand, that the issue of the *free* exploitation of information is not so important: What matters for A-consciousness is the availability of information for other cognitive processes. The idea is that if the behavior of an organism displays the presence of some sort of awareness of environmental events, then there is no reason to deny it A-consciousness.[2] Take for instance the case of subpersonal states of vision-for-action (Milner and Goodale 1995), in which a person exerts the ability to coordinate her movements appropriately following perceptual information. In this case, perceptual information is (automatically, so, in a way, freely) used in cognitive processes such as the control of action. On this view, this is a *real* case of subjects who are in a A-conscious but not a P-conscious state. And, of course, there are many other: (ordinary) blindsight turns out to be a real case too, and the same could be said for the neglect syndrome, where subjects are successful in giving the appropriate interpretation of a visual scene despite being consciously blind to them.[3]

Admittedly, it is hard to tell what happens in these puzzling diseases. But, it seems to me that we should reject the interpretation according to which both the visuomotor states and the dissociative syndromes are cases of A-consciousness without P-consciousness. I take this interpretation to be wrong because if we regarded these kinds of states as (in some sense) conscious, then *every* subpersonal state or process could be considered as A-conscious, making the notion of A-consciousness empty.

In other words, the requirement based on the (mere) availability of information for cognitive processes is too easily satisfied by plenty of subpersonal processes. As we saw, Block is not committed to the thesis that subpersonal visuomotor states are

[2] As Peter Carruthers' example points out, "mental states are access conscious when they are accessible to, or are having an impact upon, other systems within the agent (e.g., belief-forming systems, or planning systems, or higher-order thought systems, or linguistic reporting systems—it is obvious that access consciousness comes in a wide range of different varieties depending on which "other systems" are specified)" (Carruthers 2005, p. 13). Please note that Carruthers does not jump (at least, not explicitly) to the conclusion that subpersonal states of vision-for-action are A-conscious.

[3] Here too, Block would suggest that only a "super-neglect" syndrome, rather than real neglect, is what we need in order to talk of A-consciousness. Indeed, subjects affected by neglect also need a cue to provide the appropriate answers to the experimenter.

A-conscious, since, according to his own requirement, in order for a mental state to be A-conscious, its content must be poised for the *rational* control of action (see § 1 above), whereas sensorimotor coordination is an ability also possessed by pre-rational creatures. However, even if Block's notion of A-consciousness was perfectly good, there is no plausible sense in which a person entertaining an A-conscious mental state could be said to be conscious tout court (simply conscious).

Block was clearly sensitive to the problem of specifying requirements on the notion of A-consciousness which were narrow enough to discriminate between "pure" subpersonal processes and A- and not P-conscious processes; however, as we saw, no empirical, concrete case matches the constraint: He needed to devise the "super-syndromes." So, we are faced here with a dilemma of a familiar kind: either we are very liberal in ascribing A-consciousness, and in this case too many brain processes turn out to be conscious; or we appeal to Block's constraints, and in this case it is hard to show that there actually are cases of A-consciousness without P-consciousness, unless one is prepared to include, among the genuine empirical cases, such "semi-fictitious" cases as super-blindsight or super neglect. Since I am not, my conclusion is that there is no such thing as a non-phenomenal A-consciousness.

Note that this conclusion is also closer to the pre-theoretical intuition, to the extent that we should take note of it: The layman would hardly say that subjects affected by the neglect or blindsight syndrome are conscious. Intuitively, it is hard to see why a person in a phenomenally nonconscious state should be said to be in a conscious—simply conscious—state.

Moreover, even if we wish to grant Block that the "super-syndromes" are genuine cases of "access without phenomenology," we could still forgo the distinction between P-consciousness and A-consciousness, provided that we are able to account for the same kinds of phenomena by a single notion of consciousness, one that fits common sense more closely. This is the aim of the next section. My thought is that the notion of A-consciousness is the outcome of an abstraction process carried on to match certain theoretical goals; but if one is able to show that we do not get any particular theoretical payoff from this abstraction, then we have no reason anymore to maintain the concept of an A-conscious state.

To sum up, my argument against Block's distinction is the following: The alleged instances of A-consciousness without P-consciousness are not conscious states at all; and the alleged cases of P-consciousness without A-consciousness can be redescribed as cases of consciousness (simply consciousness, which necessarily involves a "phenomenal return"—some phenomenal effect) *without attention*. Therefore, we do not need to duplicate the notion of consciousness.

Yet, a question is still open. The mental states regarded by Block as P-conscious and not A-conscious are also somewhat puzzling: In what sense can we regard them as conscious if subjects are not *currently* aware of it? (they report they were aware of it only after a while). That is to say, is consciousness without attention consciousness enough? Or, from a slightly different point of view: If one does not seem to have currently an access to an alleged phenomenal state, how can the thesis be vindicated that this state is conscious? My proposal starts precisely from reflection on this problem: I will assume that there is no P-consciousness without some kind of access. Importantly, I mean here by "access" that the whole *person* has access to the state.

17.3 From "Pure" Phenomenal Consciousness to Ordinary Consciousness

As I noted above, Block's cases of P- and not A-consciousness can be reinterpreted as follows. A piece of sensorial information (about either the external world or our own body), supposedly processed by "low-level" systems, is not available for high-level processors, for different reasons (remember that there are both the ordinary case and the experimental case): either attention is focused elsewhere, so that the information does not get through the attention filter, or the sensorial data exceed the capacity of working memory. In both cases the relevant information cannot be processed by high-level systems—as Block would put it, it is not poised for cognitive processes (inferential or linguistic).

The crucial question is how to vindicate the claim that we are conscious or aware of something in these cases. Clearly, the availability of information to low-level processors is not a sufficient condition. Indeed, the relevant question is not whether a certain piece of information is available somewhere within the system; what we want is that the information is available to the system *as a whole*—this is, plausibly, the meaning of "being conscious." Curiously, we are faced here with the same problem we had with the alleged cases of non-phenomenal A-consciousness: We are looking for a way to draw a distinction between conscious states—conscious in the sense we are trying to characterize—and the subpersonal states postulated by cognitive science. We think that a clear statement of this distinction is mandatory, on pain of losing any plausible notion of a conscious state. Clearly, the brain/computational processes which never emerge to awareness are one thing, and the states or processes described previously that appear to be conscious to a certain extent (or in some sense) are quite another. There are cerebral states we cannot access at all, but this is not the case of the states we are discussing.

The main reason for arguing that these are bona fide conscious states is the subject's witness: Subjects report to have *seen* the letters, in Sperling experiment; you claim to have actually *heard* the noise in an ordinary situation like the one described by Block. However, it is well known that the reliability of reports must be carefully assessed. Many experiments (whose paradigm is constituted by Nisbett and Wilson 1977) have shown that persons are often or even systematically mistaken in describing the mental causes of their behavior; they confabulate and provide ex post rational reconstructions which do not match what had actually gone in their minds.

Nevertheless, it would seem preposterous to deny *any* degree of trustworthiness to what subjects report. On the one hand, it is worth pointing out that confabulation typically concerns the mental causal antecedents of behavior: Experiments have not shown that subjects were systematically wrong in reporting their conscious contents. On the other hand, subjects can better be said to be inaccurate, perhaps even *very* inaccurate, rather than plain wrong.[4]

[4] Reports are inaccurate probably because of the limits of working memory: Subjects saw some letters but they forgot them. Although the data are compatible both with this explanation and with the denial that subjects were conscious of the stimuli, the former is better. Indeed, if one endorses

Moreover, at least in the case of Sperling's experiment, even if subjects were not able to recall some details of the stimulus, they were well aware of certain aggregates of letters, that is, of Gestaltic chunks present in the stimulus. In other words, there are different degrees of *fineness* or accuracy in their access to (conscious) contents, but it is hard to put in question their consciousness of certain contents. Different contents in a given instant are potentially available to subjects, but not all the contents can simultaneously be under the focus of attention; however, small shifts of attention are sufficient to make phenomenally salient a piece of information that was not some minutes before.

This sort of "weakened," somewhat elusive consciousness suggests another interpretation, different from Block's. I think that it does not make sense to talk about a P-consciousness that is completely detached from some kind of subject's access. In the previous section, I said that non-phenomenal consciousness is not consciousness; here, I add that an alleged state of consciousness without any access by the subject is not conscious either (a similar point of view is defended by Levine 2007, p. 514). To say that a subject has a conscious experience (the specification "conscious" is indeed redundant, since the conscious character is internal to the concept of experience) *is* to say that he has some kind of access to some information, even if, in the elusive cases under discussion, the access is weak—that is, the cognitive-behavioral effects are modest or totally absent—and fundamentally passive.

Some authors express the view I am outlining by resorting to the notion of nonconceptual content. The idea, as it is stated for instance in Dretske (1997), is that there are phenomenally conscious states with a nonconceptual content, that is to say, one can be in these states without possessing the relevant concepts. In particular, in the cases discussed above, the subject does not bring to bear the relevant concepts (even if she happens to possess them). Thus, in the case of Sperling's experiment, subjects have the visual experience of all the letters, but are unable to conceptualize most of them as letters of a certain kind. Likewise, when you "hear" a noise without realizing (being aware) you are hearing it, you are entertaining a phenomenal state in which the relevant information is processed at a nonconceptual (or non-epistemic) level.

I do not believe that the notion of nonconceptual content gives us all we need to single out the class of phenomenal states we are trying to characterize, for at least two reasons. First, the notion of content is notoriously elusive, especially in the cases of perceptual experience. Second and more important, at least some authors (for instance, Bermúdez 1995) claim that even subpersonal states have a (nonconceptual) content, so we would be faced again to the problem of discriminating a certain class of conscious states from the very large class of (nonconscious) subpersonal states. By contrast, the notion of nonconceptual content is employed to draw a distinction between two other classes of mental states, which can be characterized, grosso modo, as thoughts, on the one hand, and experiences, on the other hand.

---

the latter explanation, it is hard to escape the consequence that there is no consciousness in every circumstance in which there is no memory.

I am sympathetic with this distinction, but the reader must be warned against taking the conceptual/nonconceptual distinction as a difference between two different kinds of *conscious* states. That is not my point. What I am interested to is to vindicate the conscious character of the alleged P-conscious but not A-conscious states (on Block's view). The notion of nonconceptual content could help to clarify why, but I prefer to put things in the following way.

When a subject is in the relevant kind of state, the lack of a higher-processing level makes quite fleeting and elusive the first-person effects that are joined to perceptual or somato-sensorial representations. However, as soon as higher processing is activated (that is, as soon as these representations fall again under the focus of attention), the first-person effects become palpable again. To put it in a slogan, if, on the one hand, feeling does not require thinking, on the other hand, thinking makes feeling more. Of course, this description is very approximate: Our ignorance of the perception/cognition *interface* mechanisms, if it makes sense to talk this way, makes it very hard to say anything more precise. Even the idea that the perception/cognition borders depend on the focalization of attention is to a large extent speculative although the distinction between early vision and high-level vision is characterized, inter alia, this way.

A good way to understand better the nature of these states, as I have called them, of weakened consciousness is to compare them with experiential states in nonhuman animals and infants: These are cases of feeling without knowing that one is feeling. In animals and infants there is evidence of the feeling state: An infant cries when she has a colic; a dog whines when its tail is trampled on. On the assumption, which I take to be not too committal, that these behaviors are mediated by painful states "experienced in the first person," nonhuman animals and infants can be ascribed consciousness in a certain degree. In other words, the idea is that these conscious states can be properly assessed as the first-person correlates of low and intermediate perceptual and somatosensorial processing levels.

These states are not totally non-accessible (hence, they can be distinguished from pure subpersonal states): We realize, to a different extent and in different ways, that we are in these states; likewise, a dog is conscious of (= feels) its pain when its tail is trampled on, and an infant is conscious of (= feels) her pain when she has a colic. But this access is rough and not intellectually mediated; no concepts or categories and reflection are involved here. Nevertheless, it is a kind of access. We could call it "bare-consciousness," or, maybe, "0-consciousness" (consciousness of degree 0)[5]. Focusing attention improves the degree of access, but there are different degrees at which one can "inspect" or access a "content" of consciousness. Everybody knows that different persons have different abilities of reporting and explaining their own affects and feelings. At least in some cases, differences in this ability amount to differences in the fineness of discrimination of a content: For a child, all wines seem alike—wines are all equally bitter—whereas I am able to grasp certain differences, but by no means all the differences that a wine taster is able to detect. In sum, one

[5] I prefer "bare," since "0-consciousness" could give the misleading idea that it is possible to fix exactly a minimal point of consciousness.

can make "half-conceptual" or fully conceptual discriminations that correspond to more or less "rich" states of consciousness or awareness.

In what I called above "bare-consciousness" agents feel a "first-person effect," but they do not possess or they do not bring to bear any conceptual resource in order to fathom their conscious state. Their perception of their own experience is vague and indistinct. They are not able to report anything about the experience; they can just produce the behavior "appropriate" to that experience and to the causal source of the experience, for instance to the physical injury that has caused a painful experience. Nevertheless, they can be said to have some kind of access to something that has happened to them or to their body. This access capacity can be improved and refined, thanks to interactions with the world and, most importantly, with other people (of course I am referring here to the human case), so that agents can gradually develop higher degrees of access, up to what we use to call "self-consciousness," or "sense of self."

Therefore, my proposal consists in replacing Block's neat divide between two kinds of consciousness with a fuzzy distinction among many phenomenal states that we can access in a more or less fine way. Though phenomenally different, these states are all *conscious*.

This concept of a hierarchy of access degrees fits well those theories that try to explain consciousness in evolutionary terms, both in the phylogenetic and in the ontogenetic sense (e.g., Damasio 1999). Consciousness is not an all/none matter, it is rather a matter of degree. The development of higher degrees requires the lower ones. The self, the self-conscious I, appears gradually, completing its development only in the adult, but in order to have a self-conscious I, one needs to have a proto-self, both in ontogenesis and phylogenesis. Being a proto-self (or to have a proto-self) is basically to experience feelings and raw affects.

Let me conclude with the following remark. According to Neisser (2006), Block's distinction has to be read as a distinction between conscious subjectivity and non-conscious subjectivity. This way, there is room for the intuition that, if one does not notice something, he cannot be said to be aware of that thing (as Kriegel 2004 points out in a line similar to mine, if one is totally not aware of something, how can she have a *conscious* experience?). At the same time, the importance for the subject's agency of the states called by Block "pure phenomenal states"—and that I reinterpret as "bare-conscious states"—is acknowledged.[6] Thus, according to Neisser, these states are not conscious, but are nonetheless states of a *subject,* that is, *personal* states.

Well, I have no serious objection to the idea of distinguishing subjectivity and consciousness. Clearly, for Neisser genuine consciousness must involve full awareness, that is, a sort of epistemic access. It seems to me that here we are dealing more with a terminological matter than with a substantive issue. Even if we do not want to call "conscious" Neisser's "subjective" states, still these are important for the study of consciousness, since they are the basis for the emergence of full conscious states (cf. Damasio's distinction between core end extended consciousness).

[6] Neisser points to an interesting link between these states and the notion of unconscious in psychoanalysis.

Be that as it may, my point is: (1) to account for the difference between a second-order state of awareness (for instance, to reflect on one's own painful experience) and lower-level phenomenal states (the bare feeling of pain) but (2) to deny that there is a *neat* distinction between these two kinds of states, insofar as the difference is just a matter of degree. The difference is just a difference in the "fineness" of access. I say "No phenomenology without access" (and no consciousness at all without a certain degree of phenomenology), but I could also say, in a Neisserian vein, "No subjectivity without bare consciousness." The emergence of subjectivity requires at least what I call "bare–consciousness."

To take stock, a subject can be conscious (conscious tout court, we do not need any more the qualification "phenomenally") in different degrees. There are not two (or more) *kinds* of consciousness: Consciousness is one and is distributed along a *continuum,* following the degree of phylogenetic development, of ontogenetic development and, in the case of adults, trivially as appropriate (I happen to be "hardly conscious" of some things, and "very conscious" of others).

What are the explanatory payoffs of this way of putting things? The advantages turn out to be most conspicuous if the thesis of the graded approach to consciousness is combined with some more or less reductionist account of bare-consciousness states, according to which, *these* kinds of state strongly supervene on neurophysiological states. But this is a matter for another paper. It seems to me important enough to have a unique concept of consciousness, which at least fits better our pre-theoretical intuitions about consciousness and the available empirical evidence.

Acknowledgments To Kevin, tireless organizer of philosophical research, tremendously generous supporter of young people, delightful host, brilliant philosopher (in random order). With herzlichen Dank, for the great sympathy and encouragement he has given me. I hope he will forgive me for dedicating to him this chapter, which seems not to be much concerned with his philosophical work; but I am comforted by the fact that his philosophical interests are enormously comprehensive. Greetings!

References

Bermudez JL (1995) Nonconceptual content: from perceptual experience to subpersonal computational states. Mind Lang 10(4):333–369. (Reprinted in: Gunther Y (ed) (2003) Essays on nonconceptual content. The MIT Press, Cambridge

Block N (1994) Consciousness. In: Guttenplan S (ed) A companion to the philosophy of mind. Blackwell, Oxford

Block N (1995) On a confusion about a function of consciousness. Behav Brain Sci 18(2):227–287. (Reprinted with modifications in: Block N, Flanagan O, Güzeldere G (eds) (1997) The nature of consciousness. The MIT Press, Cambridge)

Block N (2007) Consciousness, accessibility, and the mesh between psychology and neuroscience. Behav Brain Sci 30:481–499

Carruthers P (2005) Consciousness: essays from a higher-order perspective. Oxford University Press, Oxford

Crane T (2002) Consciousness, the awareness of the world and the essence of mind. Exploring consciousness. Fondazione Carlo Erba, 35–45

Damasio A (1999) The feeling of what happens. Harcourt Brace, New York
Dretske F (1997) Conscious Experience. In: Block N, Flanagan O, Güzeldere G (eds) The nature of consciousness. The MIT Press, Cambridge. (Reprinted in: Dretske F (2000) Perception, knowledge, and belief. Selected essays. Oxford University Press, Oxford)
Gozzano S (2009) La coscienza. Carocci, Roma
Kriegel U (2004) Consciousness and self-consciousness. Monist 87:185–209
Levine J (2007) Two kinds of access (Comment to Block). Behav Brain Sci 30:514–515
Milner DA, Goodale MA (1995) The visual brain in action. Oxford University Press, Oxford
Nagel T (1974) What it is like to be a bat. Philos Rev 83:435–450
Neisser JU (2006) Unconscious subjectivity. Psyche 12(3). http://psyche.cs.monash.edu.au/
Nisbett RE, Wilson TD (1977) Telling more than we can know: verbal reports on mental processes. Psychol Rev 84:231–259
Sperling G (1960) The information available in brief visual presentation. Psychol Mono Gen Appl 74:1–28
Tye M (2005) Consciousness and persons. MIT Press, Cambridge
Weiskrantz L, Warrington EK, Sanders MD, Marshall J (1974) Visual capacity in the hemianopic field following a restricted occipital ablation. Brain 97:709–728

Chapter 18
Basic Intentionality, Primitive Awareness and Awareness of Oneself

Martine Nida-Rümelin

Abstract A number of philosophers use the technical term 'subjective character' with the intention to refer to what makes an experience something like *for someone*. Three different insights about experience can be taken to intuitively and implicitly motivate the introduction of that technical term: (1) The insight that all experiences involve an experiencing individual to whom something is phenomenally given (basic intentionality), (2) the insight that in every experience the experiencing subject is nonreflectively and nonconceptually aware of having the experience in the simple sense that having the experience partially constitutes the subject's overall phenomenology (primitive awareness) and (3) the insight that experiencing involves some nonconceptual and nonreflexive awareness of the experience's basic intentionality (and thereby some sort of pre-reflexive self-consciousness). Considering the resulting three interpretations of subjective character more closely, a number of widespread assumptions about 'subjective character' turn out to be untenable. It is argued that the technical term 'subjective character' is in several ways seriously misleading and should rather be abandoned.

Keywords Prereflexive self-consciousness · Self-awareness · Mineness · Primitive awareness · Basic intentionality

18.1 Introduction

Are all puzzles about consciousness puzzles about phenomenal consciousness and all puzzles about phenomenal consciousness puzzles about what it is like to have experiences of a given kind? In recent years, an increasing number of philosophers have started to think that something important has been left out of the picture when the puzzle of consciousness has been approached in this way. The philosophical problem about consciousness is not only a problem about what it is like to have an experience but also, or primarily, about the fact that it is something like *for someone* to have the experience. This is the insight shared by an increasing number of philosophers. When I am visually presented with the blue sky in a perceptual

M. Nida-Rümelin (✉)
Université de Fribourg, Fribourg, Switzerland
e-mail: martine.nida-ruemelin@unifr.ch

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_18,

experience, then it is something like *for me* to have that experience. Following Joe Levine, Uriah Kriegel and Terrence Horgan, many philosophers have started to think that it is above all this 'being-for-me' that makes consciousness mysterious.

Here is a natural way one might put the insight: The puzzle about consciousness is *not,* or not primarily, about the nature of the content of experiences (about what is phenomenally present to the subject in a given case), nor is it a problem about alleged qualitative properties of those events we call experiences. Rather, the problem about consciousness arises mainly because there is an intriguing sense in which in every single experience there is someone, an experiencing subject, *to whom* something is given in a particular way. This simple and natural way to formulate the insight is, however, almost absent in the relevant texts. The discussion does not focus on the nature of the experiencing subject and its relation to what is phenomenally given as one might have expected. The nature of the individual *to whom* something is given is not addressed and that individual is not explicitly introduced into the theoretical framework. Rather, the philosophers at issue introduce an alleged further *property* of *experiences,* the subjective character, the mine-ness or the for-me-ness of those events we call experiences.

I believe that this framework is ill founded and leads unavoidably into a number of serious mistakes. Some of those mistakes have a long tradition. They may be found, I think, for instance, in the work of Franz Brentano. I may well be wrong about Brentano, however, and hope that the philosopher, friend and specialist about Brentano to whom this festschrift is dedicated will not hesitate to make me aware of my errors in that case.[1]

In an appendix to this chapter, the reader finds a series of citations from different authors that illustrate the way the term 'subjective character' has been introduced and used in recent years. It appears obvious that in these citations the authors 'point to' something real. They are trying to capture a 'phenomenon', or rather 'phenomena' that can hardly be doubted. In what follows, I will distinguish three interrelated 'aspects' of consciousness, these descriptions may be taken to be about. I will thereby start to develop a theoretical description of these 'aspects'. It is not my purpose, however, to develop any theoretical account of consciousness. My purpose is much more modest. I would like to attract the reader's attention to three different but interrelated 'aspects' or 'phenomena' the term 'subjective character' might be taken to refer to. In doing so, I cannot avoid proposing a language (technical terms). I am, however, trying to be as little 'theoretical' as possible in order to avoid misdescriptions that one might easily be led to when guided by some pre-established theoretical framework. It is sometimes quite obvious that we are at the limits of language when we try to capture the relevant 'aspects' of consciousness. It might well be that every attempt at an adequate description is, in a sense, bound to fail.

After having introduced the three different 'aspects' the term 'subjective character' might be taken to refer to (basic intentionality, primitive awareness, awareness

[1] For Kevin Mulligan's work about Brentano, compare Mulligan (2004) and Mulligan and Smith (1985).

of basic intentionality), I will come back to a few questions that have been raised about 'subjective character' in order to see how they can be answered in the light of these distinctions.

18.2 Basic Intentionality

A first thought that appears to be present in some uses of 'subjective character', 'mine-ness' and 'for-me-ness' is the observation that any experience requires *a subject to whom something is phenomenally given*. This is, in my view, an observation concerning the metaphysical structure of experiences. Experiences are events that involve a subject which exemplifies a very special kind of properties, which I will call *experiential properties.* To have a specific experiential property is for a subject *s* to be such that something is phenomenally given to *s*. The 'something' which is phenomenally given may not exist. When one closes one's eyes, one might be visually presented with red points flying around in groups. It is a bizarre idea that these points need to be accepted as existing entities. We 'refer' to these points only for the purpose of describing the experiential property of the experiencing subject. Doing so should not be taken to commit us to the ontological claim that these red points have to be taken seriously when we think about ontology, or so I claim. Still, even in such an extreme case, where what is phenomenally presented clearly does not exist, experience has a structure that we might call *basic intentionality*: We can distinguish between the *experience* which is an event involving a subject and *the experienced,* that which is, in that event, *what* is phenomenally present to the subject. To put it simply, in any experience there is a subject *to whom* something is phenomenally given.[2]

There are a number of important, deep and difficult issues surrounding what I just called basic intentionality. What is the nature of the subject involved in any such case? What is the nature of the 'relation' between that subject to whom something is phenomenally present and the 'object' or content of the experience which is in that way phenomenally present? Is it, strictly speaking, a relation? (I think it is not.) These issues cannot be addressed here. What I am trying to do is to express a natural thought about the nature of experience which comes to mind immediately when we think about it for a while. This natural thought clearly appears to be present and is evoked in the reader in some formulations of what so-called subjective character or mine-ness is about. We thus arrive at a first possible interpretation of what 'subjective character' of experience in the sense at issue might consist in:

> *Definition 1* ('subjective character' in the sense of basic intentionality)
> An experience has subjective character iff the experience is an event which consists in the fact that something is phenomenally given to some subject.

[2] G.E. Moore (1914) formulates this point very clearly (see § 3). In M. Nida-Rümelin (2011), I argue that phenomenal presence is non-relational across the board (not only in cases like seeing points with one's eyes closed but also in veridical perception). But this controversial issue is irrelevant here.

It has been claimed that subjective character is what makes an event a conscious event. If subjective character is defined in the sense of basic intentionality, this is a quite plausible claim. The claim then amounts to saying that whenever a subject is in a conscious state or undergoes a conscious experience something is phenomenally present *to the subject*. If 'phenomenal presence' is understood in a wide sense which does not limit the phenomenal to the sensory but allows for contents of thought, for instance, to be phenomenally present in thinking and which allows, for instance, for 'my being the author of what happens' to be phenomenally present in acting, then there is, in my view, no reason to deny the claim.[3]

18.3 Primitive Awareness

Here is another thought which is present in common descriptions of what the term 'subjective character' is intended to refer to: Whenever a subject has an experience of something, the subject is not only aware of that something but also aware of having that particular experience. So, for instance, when I look at a tree in front of the window, I am not only aware of what I see, I am, in some sense, also aware of my being under the impression of seeing a tree in front of the window. Without the latter awareness of my being under that impression I cannot be under that impression and so cannot be aware of what I see. This, I think, is a way to express the relevant intuition. The idea then, using the term of 'phenomenal presence', can be put like this: Whenever something is phenomenally present to a subject *S,* the subject has a particular kind of awareness of being phenomenally presented with *X* which is constitutive of the subject's being phenomenally presented with *X*. Using the term 'experiential properties' introduced above, one can express this idea in a simpler way: Whenever a subject has an experiential property *E,* the subject is aware of having property *E* in a way which is constitutive of having *E*. For example, when you are phenomenally presented with a specific pain in your head, then you must, in a sense, *be aware of your being presented with that pain* in order for it to be true that the pain is phenomenally presented to you. I share this intuition.

Another way to put the same point is to say this: You cannot have an experience without being conscious of the experience since only conscious experiences are experiences. The problem now is to get a grip on the kind of awareness or consciousness at issue here. Let us suppose that we agree on this: having an experiential property requires being aware of having it; having an experience requires being conscious of having the experience. In this agreement, we have a specific kind of awareness or consciousness in mind. But what kind of awareness or consciousness is it? Let me call the kind of awareness we are searching for 'primitive awareness'.

[3] The claim I am here agreeing to should not, however, be confused with another: Whenever a subject is consciously aware of *x, x* is phenomenally present to the subject. This claim, I believe, is mistaken. I will come back to it in Sect.18. 10.

The preceding remarks motivate the following constraint on any account of primitive awareness:

> C1: To have an experiential property essentially involves being primitively aware of having that experiential property.

The locution 'having *P* essentially involves *Q*' means that having *Q* is constitutive of having *P*. In other words: having *Q* belongs to what it *is* for something to have *P*. C1, I believe, is an insight that helps to resolve certain puzzles about so-called transparency of experience and to avoid related confusions. It has been urged that we cannot get aware of the phenomenal character of our experience because whenever we try to gain this awareness we unavoidably focus on what is given in the experience. When we appreciate C1, we can see in what way this argument goes wrong. To be phenomenally presented with, for instance, a reddish blue, essentially involves being aware of being presented with that colour. No wonder then that focusing on the content of the experience *is* a way to focus on what one is aware of in primitive awareness: one's own having a certain kind of experience. I can focus on my having a specific kind of colour experience by focusing on the colour given in that experience.[4] C1 helps us to understand how that can be the case.

Primitive awareness is 'primitive' in the sense that it is not to be confused with other more demanding ways of being aware of one's own experiences. I may, for instance, look on a particular patch of blue and wonder whether I see a pure blue or whether there is an almost unnoticeable tiny reddishness to be discovered in the blue I am presented with. I may then realize that the blue actually has such a tiny reddish component. I then realize that I am having an experience of reddish blue. In this case, I apply a concept of a certain type of experiences which I acquired earlier on the basis of other colour experiences. Primitive awareness does not involve such a judgment. This follows from the constraint formulated above. Judging that I am presented with a reddish blue is *not* necessary for being presented with reddish blue. (A subject can have that kind of experience before the subject has acquired the conceptual capacities that enable it to judge that it is having an experience of that kind.) Furthermore, being presented with reddish blue does not require that I entertain any thought about my having that experience. I may have an experience of a particular kind without thinking about having it. So, being primitively aware of having an experiential property is somehow simpler, more primitive, than other forms of awareness one can have of one's own experiential properties. This observation motivates the choice of the term 'primitive awareness' for *the* kind of awareness which renders the constraint C1 true.

What is it to be primitively aware of having a certain experiential property? In other words: Which kind of awareness of an experiential property *E* is essentially involved in having an experiential property *E*? What is primitive awareness? Here is a proposal: A subject is necessarily aware of being presented with something in the sense that being so presented with something makes a difference for the subject at that moment. This is to say, or so I propose, the following: Having the property

[4] This idea is developed in Nida-Rümelin (2007).

at issue partially constitutes the overall phenomenology of the subject's present state. There are many properties we have at a given moment which are not constitutive of the phenomenology of our present state. Having a certain weight is one of them. Experiential properties are such that having them does partially constitute the phenomenology of the relevant overall phenomenology. My simple proposal is this: A subject is primitively aware of having *E* just in case that having *E* makes a difference for the subject: *Having E partially constitutes the subject's overall phenomenology*. Using this notion of primitive awareness, one could define a notion of 'subjective character' in this way:

> *Definition 2* (subjective character in the sense of primitive awareness):
> An experience has subjective character iff in having the experience the subject is primitively aware of having the experience.

18.4 A Few More Remarks About Primitive Awareness

It may be objected that the account of primitive awareness just given is unhelpful because it is totally uninformative. To say of an individual that it has a given experiential property (e.g. the property of being visually presented with a tree) already involves that being so presented with a tree partially constitutes how it is for that individual to be in its present state. So, if being primitively aware of having that kind of visual experience is nothing but being in a state the phenomenology of which is partially constituted by having that visual experience, then saying that *x* is primitively aware of having an experience does not add anything to just saying that *x* has that experience.

It is true that nothing is added when after having said '*x* is phenomenally presented with *y*' we add 'and, furthermore, *x* is also primitively aware of being phenomenally presented with *y*'. But this can hardly be an objection. After all, we were searching for something that cannot possibly lack when someone has an experiential property. So we are surely not searching for a condition, which when added to the claim '*x* is phenomenally presented with *y*' expresses any additional information. This is precisely required by the constraint formulated above for any account of what it is to be primitively aware of having an experience. Attributing primitive awareness does not involve attributing any *further* property to the subject, any property that has not already been attributed in attributing the experience itself.

But then, if being aware of having an experience is nothing but having the experience, what is the point of introducing, in addition to our notion of experiences or experiential properties, the notion of being primitively aware of experiences? Do we need such an additional notion? What 'work' could it possibly do?

I suggest that the notion of primitive awareness is not needed in the following sense: It is not one of those notions that must be included in the terminology used in any adequate and complete theoretical account of consciousness. The notion of a subject, and the notion of phenomenal presence or of basic intentionality belong to these central notions (I believe); the notion of primitive awareness probably not.

But the notion of primitive awareness can nonetheless be useful if it is able to capture a shared intuitive idea and if it can help to see interrelations between different implicit notions we use in our thought about issues surrounding consciousness. It seems to me that, for purposes of this kind, the notion of primitive awareness has a role to play.

The notion of primitive awareness can help to understand the special and puzzling nature of experiential properties. It has been said and written by several philosophers in the context of 'subjective' character that experiences are not only 'in me' but also '*for me*'. This formulation is a bit puzzling since it is unclear in what sense experiences are 'in me'. Experiences are, I take it, events involving an experiencing subject. They consist in the fact that something is phenomenally present to someone. Understood in this way, experiences cannot be or happen within a subject's body (or so I would like to insist). Nonetheless, it is not difficult to see what people have in mind, when they distinguish between an experience being 'in me' and 'for me'. Here is a different way to put it: Experiential properties are properties that a person not only simply has (like the property of being in Fribourg) but they are properties such that *having them* is, by itself, something like *for* the person. If this is the thought associated with the kind of awareness we are searching for, then primitive awareness is an excellent candidate. Experiential properties have the remarkable feature that having them *is* being aware of having them. And being aware of having them means that having them partially constitutes one's present phenomenology.

Another objection against my proposal is to doubt that primitive awareness is a kind of awareness. But I do think that talking of awareness here is quite adequate. The locution '*x* is aware of *y*' involves the idea that, in some way, *y* is present to *x* and that it is present to *x* in a way which makes it in principle available for *x* for reflection, conceptualization and judgment. All this, I suggest, applies to primitive awareness. When I am visually presented with a tree then my being visually presented with a tree is itself *present* to me; it is present to me in the most primitive sense: It contributes to how it is for me to be in my present state. Furthermore, my being visually presented with a tree is—under normal circumstances—available to me for reflection and judgment. On the basis of being so visually presented I cannot only judge, under normal circumstances, that there is a tree, I can also judge that I am now having a certain phenomenal kind of experience (the one which is characterized by being visually under the impression that there is a tree). In that sense, it is correct to say (in accordance with self-representationalists about consciousness) that the experience has a double function. It makes me aware of the tree and it makes me aware of my having a specific kind of experience: On the basis of the experience, I can normally judge that there is a tree *and* that it so appears to me.

One can also realize that primitive awareness is a kind of awareness in the following way. It is obvious that a person who is conceptually equipped for the relevant kind of judgments can judge that he or she is under the impression of there being a tree simply by having the experience (by being under the impression). If this is so, then he or she must have been aware of being under the impression in having the experience. I here apply a plausible principle: One can judge that *p* simply by having a particular experience only if in having the experience one is aware of *p*.

But then the question arises: In what sense is the person aware of being under a certain impression by being under that impression? The answer here suggested is: He or she is aware of being under the impression of there being a tree in the sense of primitive awareness and this grounds his or her judgment that he or she is under that impression.

In the debate about so-called transparency, it has often been said that when you try to focus on the subjective character of your experience, then you unavoidably find yourself focused on the content of your experience (on what you are presented with in having the experience). For instance, when you try to focus on your being under the visual impression of there being a tree, you focus on the apparent tree 'out there'. This observation has often been taken to motivate the view that we cannot focus on an experience (in other words: one one's being under a certain impression). But this is, I think, a mistake. A subject may focus on its being under the impression of there being a certain object with certain properties by focusing on the apparent properties of the apparent object. There is no opposition (or there need not be an opposition) between these 'two' acts of attention. Is it even inadequate to talk of *two* acts of attention here at all?[5] The proposal just given might be used to explain why there is only one act of attention involved. When I focus on what I am aware of by being visually under the impression of there being a tree, then I focus on the apparent tree; and when I focus on what I am aware of in my primitive awareness of being under that impression, then I focus on how things appear to me, on what it is like for me to be under that impression. How can I do both at once? Here is a tentative answer: having the experiential property of being under the visual impression that there is a tree involves (a) visual awareness of the apparent tree and (b) primitive awareness of being under that impression. Therefore, focusing on having an experiential property *is* then focusing on two objects one is both aware of in having the experiential property. Focusing one's attention on one's having the experiential property of seeing a tree involves attending to the tree (to what one is perceptually aware of in having the experience) *and* to one's being under the impression of there being a tree (to what one is primitively aware of in having the experience).

18.5 Awareness of Basic Intentionality

According to many philosophers belonging to different traditions, there is a sense in which consciousness necessarily involves some simple nonconceptual form of self-consciousness which does not require any kind of reflection upon oneself or upon one's own mental states.[6] This is an idea sometimes referred as 'pre-reflexive

[5] For discussions of transparence of perceptual experience compare, for instance, Tim Crane (2002), Fabian Dorsch (2011), Amy Kind (2003) and (2010), Mike Martin (2002), Martine Nida-Rümelin (2007), Charles Siewert (2004), Daniel Stoljar (2004) and Michael Tye (2002).

[6] An idea of this kind plays a role, for instance, in Husserl (1900/1901/1984) and Sartre (1936) and, more recently, I believe, also in Cramer (1974), Frank (2011), Henrich (1970) and Pothast

self-consciousness'. This idea appears to be involved as well when contemporary philosophers use the term 'mine-ness'. To get some clear understanding of this possible further interpretation, it is necessary to reflect upon the following questions: *Is it true that we are aware of ourselves in some sense in every experience?* Is there some primitive form of self-awareness which is necessarily involved in any conscious episode? More precisely: Is there a sense in which every subject who undergoes a conscious experience is thereby necessarily aware of itself?

According to the view I would like to propose, all these questions must be answered in the positive. There is a sense in which one cannot consciously experience anything without thereby being aware of oneself in a nonconceptual and nonreflexive manner. I would now like to propose a way in which one might describe that fundamental and omnipresent kind of self-awareness.

At the beginning of this chapter, I made the following proposal about the metaphysical structure of experience: Every conscious experience exhibits basic intentionality; it consists in there being a subject *to whom* something is phenomenally given. My proposal, now, is this: In having an experience we are necessarily aware of that structure. We are aware of that structure by being the subject involved in the experience, by being the one to whom something is given. Furthermore, by being aware of that structure, by being aware, in every experience, of its basic intentionality, we are aware of ourselves as the one to whom something is phenomenally given. According to this proposal, the relevant kind of awareness of oneself does not give access to some object, oneself, in isolation. To say it in a metaphorical manner: This kind of awareness of oneself does not require turning 'the mind's eye' towards oneself or 'back' to oneself. It would be misleading and inadequate to draw a picture here with two arrows, one directed at what is experienced, the other directed back to oneself. One arrow directed towards what is given is sufficient to pictorially represent the real situation. Awareness of basic intentionality, as it is present in every experience, does not require a concept of oneself or a general concept of what it is to be a subject of experience. It is there before we start to reflect upon what is given in the experience or upon the question about who (or what kind of individual) it is to whom something is given. That kind of awareness is there, not on the basis of experiencing, but in the act of experiencing itself.

That there is this awareness of basic intentionality may become more plausible if one reflects upon why and how one comes to agree with the thesis about basic intentionality proposed earlier. Those who agree with the claim are likely to agree, furthermore, that the claim is true in an obvious manner. When one accepts the thesis, one will accept it immediately without any cognitive effort. One does not

(1971). Awareness of basic intentionality (and perhaps sometimes primitive awareness) might be a way to interpret or might at least be helpful to clarify what different authors have in mind when they talk of pre-reflexive self-consciousness (some consciousness of oneself that is necessarily present in every conscious state and which does not require any kind of reflexion). It would, however, require careful examination to decide to what extent this interpretation can capture what different authors sharing the basic idea of 'pre-reflexive self-consciousness' have in mind. I will leave this question open here. An excellent exposition and discussion of views about pre-reflexive self-consciousness can be found in Shaun Gallagher and Dan Zahavi (2005/2010).

need to carefully examine general features of experience on the basis of memory or imagination. Why is the thesis so obvious? The claim that we are aware in every experience of its basic intentionality provides an answer. We are aware of that fundamental structure all the time in our conscious life; that is why it is so easy to agree that all experiences have that structure once we have formed the relevant notion.

The preceding remark, however, can easily be misunderstood. It might appear as if the proposal here is to infer the existence of some pre-reflexive and nonconceptual awareness of basic intentionality from our intuitive tendency to agree with the relevant theoretical claim. But this is quite contrary to my intention. The thesis that we are aware of basic intentionality would be poorly supported if it were in need of such an indirect and theoretical argument; and such an indirect argument could not help much to improve its plausibility if we did not have a more direct and immediate access to it. A far more important way to come to agree with the present thesis is by reflection upon how it is to experience something. One can realize, on that basis, or so I claim, that there is a sense in which we are unavoidably aware of an experience's basic intentionality in undergoing it. The claim, to say it with respect to a concrete example, is the following. When a tree is phenomenally given to you, you are, in being under that impression, aware of there being something that is phenomenally present to you. This kind of awareness is not something the experience of the tree might have or lack; it is rather part of what it is to have an experience of there being a tree. According to this proposal, having an experience necessarily involves awareness of there being something that is 'given to me'; the structure of the experience, its basic intentionality is evident to the experiencer, or 'reveals itself' to the experiencer in having the experience.

We thus arrive at a further sense in which an experience may be said to have 'subjective character':

> *Definition 3*: (subjective character in the sense of awareness of basic intentionality)
> An experience has subjective character iff the experiencer is aware, in undergoing the experience, of its basic intentionality.

18.6 A Few Remarks About Awareness of Basic Intentionality

Is awareness of basic intentionality itself an experience with the structure of basic intentionality? A positive answer to this would lead into trouble. We then would have to say that in being aware of an experience's basic intentionality we again have an experience with basic intentionality of which we would have to be aware, and so on. This regress problem is a reason to answer the above question in the negative. But it should not be the only reason; otherwise this negative answer would be too ad hoc. There should be an independent reason to say that awareness of basic intentionality in having an experience is not itself an experience which exhibits basic intentionality.

And there is, or so I propose, an independent reason for this claim. If basic intentionality were something we experience then basic intentionality should be phenomenally given in having an experience. This would mean that, when you see a tree, there is a tree phenomenally present to you and, in addition, there is something more phenomenally present to you, namely the metaphysical structure of your experience, its basic intentionality. But this is a misdescription. We are aware of basic intentionality in every experience in a way which does not add anything to the content of the experience, it does not add anything, in other words, to what is 'in front of the mind's eye', to the totality of what is phenomenally present to the subject. The metaphor of the stream of consciousness might help to make this point a bit clearer. The stream of consciousness is the totality of what is phenomenally given, it is an extremely complex and rich totality of what is given to a subject in perception, emotion, bodily feeling, memory, imagination and thought, a totality which is in permanent change from moment to moment. The stream of consciousness, so understood, does not 'contain' the subject, it is rather the totality of what is present to the subject over a stretch of time. To say that basic intentionality is not phenomenally present, or to say that awareness of basic intentionality is not a kind of experiencing is to say—within the metaphor of the stream of consciousness—that basic intentionality does not occur in the stream of consciousness, it is not an element in it, among others. Rather, we are permanently aware of the basic intentionality of experience in experiencing, in being presented with the rich totality which makes up the stream of consciousness within a given period of time. We should not think of the necessary relation between phenomenal consciousness and awareness of basic intentionality as relating two phenomena with one another. Rather, awareness of basic intentionality is an aspect of what it is to be phenomenally conscious of something. This is why basic intentionality does not enter the content of the experience. We are not aware of basic intentionality by experiencing it as a further element in what is phenomenally 'there'.

Awareness of basic intentionality is—as mentioned earlier—a form of self-awareness. In being aware of the metaphysical structure of experience in undergoing an experience we are aware of the one to whom something is phenomenally given, we are aware of ourselves, or so I suggest. It has often been observed that the subject is not given to itself as an object; different versions of that claim and different aspects of the phenomenon it is about have been a central theme of important work within the philosophy of consciousness. I cannot discuss these deep issues here but I would like to add the following observation. If we understand self-awareness—in the fundamental sense in which it is included in every conscious experience—as awareness of basic intentionality, then we can see why and in what way the subject is not given 'as an object' in that kind of self-awareness. Basic intentionality is not an element in the stream of consciousness, and this is why 'the self' does not occur either in the stream of consciousness when a subject is self-aware in being aware of his or her experience's basic intentionality.

It is quite common among philosophers to think that 'the self', the experiencing subject, somehow comes into existence by gaining self-consciousness. This is, in my view, a mistake. The simplest experience already has the structure of basic in-

tentionality and thereby presupposes for its existence the existence of a subject. The subject of the experience is not created by the acquisition of self-consciousness. Using what has been said about basic intentionality and awareness of it, there is, however, a way to see why this is a mistake one can easily make. The idea that 'the self' is created by the subject's acquisition of self-consciousness has no initial plausibility if 'self-consciousness' is understood in a conceptually demanding way. But it may have a prima facie plausibility if 'self-consciousness' is understood in the sense of pre-reflexive self-awareness which I here call 'awareness of basic intentionality'. To think that the subject comes into being due to its acquisition of pre-reflexive self-awareness in this sense is a mistake we can now explain. To do this let us assume, for the sake of argument, that the experiencing subject comes into being when, for the first time, it starts experiencing. If we assume, furthermore, that any experiencing necessarily involves awareness of basic intentionality, then the subject comes into being when it gains, for the first time, this form of self-consciousness (awareness of basic intentionality). If this is so, then, necessarily, a subject comes into being when it acquires, for the first time, this kind of self-awareness. But this kind of self-awareness does not constitute the subject's existence and it does not bring it about. Given the necessary simultaneity just mentioned one can, however, quite easily be tempted to draw this mistaken conclusion. One can see, however, that the conclusion is mistaken by appreciating that basic intentionality does not come about by the subject's awareness of it. To think so is to confuse awareness of something with 'the something' one is aware of in that awareness.

'Mine-ness' or 'subjective character' is used, or so I claim, in a systematically ambiguous way, sometimes as referring to basic intentionality (first reading, definition 1), sometimes as referring to primitive awareness (second reading, definition 2) and sometimes as referring to the subject's awareness of basic intentionality (third reading, definition 3). If one overlooks the ambiguity of 'subjective character' between its first and its third reading then one commits exactly the mistake just discussed: One does not realize that basic intentionality cannot be equated with awareness of basic intentionality and thereby implicitly accepts a view involving the claim that 'the self' is constituted by some form of pre-reflexive self-awareness.

I said above that awareness of basic intentionality is not a case of experiencing. In other words: Basic intentionality is not among the elements that are, in the experience, phenomenally given to the subject. This claim might easily be confused with a different one: with the claim that awareness of basic intentionality is no part of what it is like to experience. The latter claim, I suggest, is false. It is a *phenomenological* fact about *all* experiences that experiencing necessarily goes along with awareness of basic intentionality. This is, however, a special phenomenological feature: It is not a feature, as in the normal case of phenomenal features, which distinguishes one experience from another. It is rather an aspect of how it is to experience shared by all experiences. With this clarification in mind we can describe the above mistake in yet another way. To confuse basic intentionality (first reading of 'mine-ness') with awareness of basic intentionality (third reading of 'mine-ness') is to confuse metaphysics with phenomenology. Again, it is easy to make the mistake given this

unusual fact about basic intentionality: Its occurrence necessarily involves a particular phenomenal feature, awareness of basic intentionality. Furthermore, our best reason to accept that experiences exhibit basic intentionality is this particular phenomenal feature shared by all experiences (our awareness of basic intentionality present in every experience). As mentioned earlier, acceptance of that thesis about the metaphysical structure of experience is based on our omnipresent awareness of that metaphysical structure and it is based on it in an immediate way which does not require any conscious inferences. There is, however, a belief with a different content which is based in a similar direct way on that awareness: the belief that we are so aware of that structure in any experience. Given their quite similar relation to our awareness of basic intentionality involved in any experience, one might confuse the two beliefs and thereby fail to distinguish their different contents: the content of the first is a metaphysical fact, the content of the second is an experiential fact. The experiential fact is that, in experiencing, we are aware, in a phenomenally relevant way, of the metaphysical fact at issue (of basic intentionality).

18.7 Subjective Character and Phenomenal Character

It is often said that experiences have qualitative character *and* subjective character. A naive way to read this which appears to be quite widespread is to think of both as parts of what-it-is-like to have an experience. The qualitative character, according to a specific way to understand that naïve reading in the special case of perception, is characterized by how things appear to be in that experience or by how they appear in that experience.[7] The subjective character of the experience then must be a further fact about what it is like to have an experience which is not exhausted by its qualitative character so understood. According to my diagnosis, there are three interpretations of 'subjective character'. The question about whether subjective character is part of phenomenal character therefore divides into three.

In the preceding section, I argued that subjective character in the sense of definition 3 is indeed part of the phenomenal character of experiencing. It is a fact about phenomenology, about what it is like to experience. However, it is not a feature of experiencing that could be absent in any possible experience, or so I claim. Experiencing necessarily goes along with awareness of its basic intentionality. For that reason, awareness of basic intentionality can well go unnoticed. We cannot discover that phenomenal feature by contrasting what it is like to have one experience with what it is like to have another experience. In order to discover that feature, we have to abstract from all specific features of experiences, we have to abstract from what is phenomenally given in a particular experience. The intellectual activity one has to engage in to discover that feature is therefore quite different from other cases of phenomenological reflection. Nonetheless, or so I would like to insist, a complete

[7] Afterimages appear red but they do not appear to be red, or so one might plausibly say. Compare for this issue Crane (2002, p. 8) and Boghossian and Velleman (1989/1997, 91/121).

description of what it is like to have a particular experience would have to mention pre-reflexive nonconceptual awareness of basic intentionality (which I take to be a form of self-awareness). Presupposing that we know we are talking about an experience, describing it as one that involves awareness of basic intentionality is totally uninformative. It does not add any information under that presupposition since the phenomenal feature described is necessarily shared by all cases of experiencing; the description does not capture any feature distinctive of the particular case at issue. Being aware of basic intentionality characterizes the overall phenomenology of any experience—even of the simplest experience lived through by some 'lower' animal. This is so although it would be misleading to say that awareness of basic intentionality *makes a difference* for the subject's overall phenomenological state. That it makes a difference might be read as saying that the overall phenomenological state of the subject *would be* phenomenally different if it lacked that awareness. But this reading makes the claim paradoxical: The subject would not be in any overall phenomenological state at all if it lacked awareness of basic intentionality.

As mentioned before, contrary to subjective character in the sense of definition 3, subjective character in the sense of definition 1 (basic intentionality) is not a phenomenological fact and so is not part of the phenomenal character of the experience, or so I claim. Basic intentionality is a metaphysical fact concerning the nature of experiences. It is not a fact about what it is like to experience; it is not constituted, as motivated above, by any experiential fact. However, subjective character in the sense of basic intentionality is revealed in every case of experiencing to the subject concerned. It is revealed in the sense that the subject is nonconceptually and prereflexively aware of that metaphysical structure; it is revealed by how it is to experience. It is, however, revealed in a way which does not include cognitive access to that fact. One has it without attending to it; a subject can have it without ever having formed a belief about the structure of experience; no concept usable in a belief about it or in entertaining a thought about it is required for having that awareness.

The third question to ask is about primitive awareness. Is primitive awareness an instance of phenomenal consciousness? In other words: When you are primitively aware of seeing a tree, are you thereby in a state which is itself part of phenomenal consciousness? In yet other words: Should we say that being primitively aware of seeing something is a case of being phenomenally aware of one's own experience? Or one may put the question in this manner: Is one's primitive awareness of having some given experiential property itself an experiential property?

The first two questions may be taken to be synonymous without any danger of confusion, as far as I can see; the third, however, must be distinguished from the former two. The first two questions can be formulated more precisely by explicating a positive answer in this way: Being primitively aware of having a certain experience characterizes the overall phenomenology of a subject's experience. In a sense, this affirmation is trivially true. Being primitively aware of having property *F* means, according to the definition proposed, that *F* partially constitutes the subject's overall phenomenology. If this is a fact about *F* (that it partially constitutes the subjects present phenomenology) in a given case, then it is a fact about the subject's overall phenomenology in that case and so to formulate that fact is to characterize the

subject's overall phenomenology. However, that this fact partially constitutes the subject's phenomenology does not mean that being primitively aware of having a certain experiential property is itself a further experiential property distinct from the one the subject is primitively aware of. In seeing a tree, you have the experiential property of being so 'appeared to'. In seeing the tree, you are also primitively aware of seeing the tree: Your overall phenomenology is partially constituted by the fact that you are under that visual impression and, in this sense, your perceptual experience is conscious. But this does not mean that you experience your own experience of the tree, in other words: It does not mean that there is any sense in which you experience yourself as being visually presented with a tree. There is no experience having your experience of the tree as its content. Being primitively aware of seeing a tree does not consist in having an experience which is itself the content of the experience. So we must conclude: Being primitively aware of having an experiential property is *not* an experiential property.

18.8 Peripheral Inner Awareness

According to Uriah Kriegel, the subjective character of experiences consists in the subject's peripheral inner awareness of the experience at issue (Kriegel 2005, 2009; Horgan and Kriegel 2007). According to this thesis, when you are perceptually aware of a tree, you are thereby, in a peripheral way, also aware of that perceptual experience. This further awareness, your awareness of the experience, is peripheral in a way comparable to the way in which one is only peripherally aware of an object in the periphery of the visual field, or in the way one is only peripherally aware of the melody played by the violas in an orchestra as long as one does not attend to them. I find this way to describe the situation at least misleading. To say why, it will help to address, for all three senses of 'subjective character' distinguished earlier, the following question: Is subjective character of an experience a case of peripheral awareness? In other words: Is 'the experience *E* has subjective character' always or sometimes true in virtue of the fact that the experiencing subject at issue is peripherally aware of something?

If subjective character is understood in the sense of basic intentionality (definition 1), then these two questions must be answered in the negative. Basic intentionality is not a fact about phenomenology, it is a fact about the metaphysical structure of experience. Therefore, experiences do not have subjective character in that sense in virtue of the subject's awareness or peripheral awareness of the experience *E* or of anything else.

If subjective character is understood in the sense of primitive awareness (definition 2) then these two questions must be answered in the negative too. For any kind of awareness to allow for the distinction between focal and peripheral awareness that kind of awareness must consist in some content being present to the subject such that that content can occupy a more or less central position among all that is given to the subject in the relevant moment. Primitive awareness, in other words,

would have to be a case of basic intentionality. In other words, the experience the subject is aware of would have to occur, among other items, in the stream of consciousness. But primitive awareness is not of that kind. To be primitively aware of one's own experience is not to have an experience of one's experience. The experience is not presented to the subject in its primitive awareness of the experience, rather, the subject is primitively aware of experiencing just by experiencing, and just in virtue of the fact that its overall phenomenal state is determined (partially constituted) by its having the experience. Therefore, or so I conclude, primitive awareness does not allow for the distinction between 'focal' and 'peripheral'.

But there certainly is a sense in which a person can be more or less centrally aware of how it is for him or her to have a given experience. A painter may be acutely aware of the colour in which something appears to him or her and thereby of the subjective character of his or her experience while another person, exposed to the same scene, might well see the objects in exactly the same colours and yet barely notice their particular quality. It would however be a mistake to think that this is a difference on the level of primitive awareness. It is a difference on the level of reflection upon one's experience. The painter is interested in how things appear to him and he therefore directs his or her attention correspondingly. His or her reflective awareness of how things appear to him or her plays a central role in his or her overall state. Another person might have no reflective awareness of the colours he or she is experiencing or he or she might reflect on these colours but in a much less focused manner while other things, more important ones, pass by 'in his or her mind'. The plausible thought that we can be more or less focally aware of our own experience should be understood, or so I propose, as a thought about reflective awareness of one's own experience. It must not be understood as a thought about primitive awareness.

The remaining question about whether subjective character may consist in some kind of peripheral awareness concerns awareness of basic intentionality (subjective character in the sense of definition 3). The question to ask then is this: Does awareness of basic intentionality allow for the distinction between focal and peripheral awareness? Is this kind of awareness sometimes more focal and sometimes more peripheral? The answer depends, again, on whether or not awareness of basic intentionality has itself the structure of basic intentionality. Awareness of basic intentionality allows for the distinction at issue only if in that awareness something is present to the subject among other objects occurring in the stream of consciousness; but, as noted earlier, this is not so. Awareness of basic intentionality is not experiential in that sense, it is not a case of being phenomenally presented with something. I conclude that the distinction 'focal' versus 'peripheral' does not apply to awareness of basic intentionality; it does not apply to that kind of pre-reflexive nonconceptual self-awareness.

As before, one might object that there is a sense in which awareness of basic intentionality can be more or less central in the subject's overall phenomenology. I agree that this is so. When one directs one's attention in a particular way then that kind of awareness might be very clearly 'in front of the mind's eye' and it may then be central in some sense in one's overall phenomenology. In this case, however, or

so I suggest, one makes one's own awareness of basic intentionality the object of one's thought and that thought might be more or less central. So, once again, the conclusion we should draw is this: The distinction 'focal' versus 'peripheral' applies on the second level, on the level of reflection, it applies to our thoughts about awareness of basic intentionality; but it does not apply on the first level, on the level of awareness of basic intentionality itself. In thinking about one's awareness of basic intentionality, basic intentionality is present in one's thought. In that thought, one is also aware of basic intentionality, but this is not awareness of basic intentionality in the sense in which the notion has been introduced here: it is not pre-reflexive nonconceptual awareness of basic intentionality. One's cognitive awareness of basic intentionality when we reflect upon pre-reflexive nonconceptual awareness of basic intentionality does, however, allow for the distinction between 'focal' and 'peripheral'.

18.9 The Regress Problem

Proponents of self-representationalism about consciousness claim that every experience has a double function: It makes the subject aware of some object (e.g. a particular visually given scene) and it makes the subject aware of having the experience (or, as they say, of the experience itself). For the representationalist, the following premise is common ground: An experience makes a subject aware of some *x* if and only if it represents *x*; to render the subject concerned aware of *x* is, for an experience, to represent *x*. Contemporary self-representationalists cite Brentano as a traditional source of this idea and they are right in doing so. In various places, Brentano clearly expresses the idea that conscious experiences involve a double representation ('Vorstellung'): in hearing a tone, the tone is represented but the hearing is represented as well.[8]

Contemporary self-representationalists share with Brentano an important insight: It would be a mistake to think that the relevant awareness of the awareness of the tone is a further act of consciousness or 'a further experience'; they are eager to stress that we must describe the situation in a way which avoids the introduction of various levels and they cite two reasons: (a) it would be phenomenologically inadequate to introduce more than one level in that description (a subject is aware of his or her experience by having the experience and without thereby undergoing a further experience) and (b) the alternative (the introduction of a higher level) leads into a serious regress problem (for the relevant act of consciousness at the second level to be conscious there would have to be an act at the third level directed at the one on the second level, and so on).

Brentano and contemporary self-representationalists try to do duty to phenomenology and try to stop the regress problem by the identification of the act of consciousness or experience in which the act of consciousness (or experience) is represented with that

[8] Compare citation (12) in the appendix to this chapter.

represented act or experience itself. But there is reason to doubt that they thereby solve the regress problem while doing duty to phenomenology. The problem, in my view, is their use of one single notion of awareness (or representation) in that context. This, I believe, leads them into trouble.

Independently of how experiences are counted, the following problem remains. If in every experience the subject must be aware of having the experience and if being aware of having an experience is itself a case of experiencing, then there is an infinite number of 'being aware of' involved: the subject is aware of being presented with *X,* the subject is aware of being aware of being presented with *X* and the subject is aware of being aware of being aware of,…, etc. *This* regress does not stop through the identification of all experiences involved at each level with one another.[9] The regress occurs because awareness is understood—across the board—as experiential awareness. The philosophers at issue seem to believe that we are talking about awareness in the same sense when we say that the subject is aware of the tone and aware of hearing the tone: the tone is given to the subject and, in addition, the experience is given to the subject. If this were so, then the second given-ness for it to be conscious would require, again, a further instance of awareness, and so on. The solution, I suggest is to realize that awareness at the first level *is* a case of basic intentionality which does necessarily involve awareness of experiencing in the sense of definition 1 and 2, but that the latter items of awareness (primitive awareness and awareness of basic intentionality) are *not* cases of basic intentionality and therefore do not necessarily involve any further awareness.

Before I explain how the regress stops in the case of primitive awareness (definition 2) and in the case of awareness of basic intentionality (definition 3), let me briefly come back to 'subjective character' in the sense of definition 1 (basic intentionality). Subjective character is commonly explained by the self-representationalist as the property of the experience to represent itself. This is, however, trivially, a nonstarter if subjective character is understood in the sense of basic intentionality. Basic intentionality is a fact about the metaphysical structure of experiences and it is not constituted by any kind of awareness. Since the experience represents itself, according to the representationalist, if and only if it makes the subject aware of the experience itself, subjective character is reduced by the self-representationalist theory to some kind of awareness. Basic intentionality is not a kind of awareness; therefore, it cannot be so reduced. I conclude that self-representationalism fails as a theory about the nature of consciousness since it cannot explain basic intentionality.

The regress problem does not arise if the way in which, in any experience, we are necessarily aware of experiencing is understood in the sense of primitive awareness. The regress starts only if we must reapply the principle at issue (any experience involves an awareness of itself) to the relevant awareness of the experience. But the

[9] I am indebted to Emmanuel Baierlé for having attracted my attention to this problem in a seminar discussion. According to Zahavi (2006), Aron Gurwitsch was the first author who clearly formulated it as an objection against Brentano in his habilitation which was completed in 1931 and published in 1977 (Gurwitsch 1977). Zahavi (2006) attributes the same objection against Brentano to Dieter Henrich (1970), Konrad Cramer (1974) and Ulrich Pothast (1971).

principle does not apply to primitive awareness because primitive awareness, as explained earlier, is not itself an experience. The same remark is true about awareness of basic intentionality. This kind of pre-reflexive nonconceptual self-awareness is not an experience either, it is present in every experience but we do not thereby experience basic intentionality, basic intentionality does not occur among the objects present to the subject in having the experience, or so I claim. Therefore, we need not and should not reapply the above-mentioned principle and therefore no regress problem arises either if subjective character is understood in the sense of awareness of basic intentionality.

18.10 The Representationalist Mistake

Primitive awareness and pre-reflexive nonconceptual self-awareness in the sense here proposed (awareness of basic intentionality) are both puzzling kinds of awareness. Language leads us into trouble when we wish to adequately describe their nature. We wish to say and we can hardly avoid saying that cases of primitive awareness are cases where a subject is aware *of* having an experience, and we cannot avoid the 'of-talk' either when we talk about the aspect of consciousness at issue calling it 'awareness *of* basic intentionality'. In the first case, we easily slip into saying that the subject is aware *of* the experience. But this sounds as if there is something, the experience, which is, in such a case of primitive awareness, present to the subject or given to the subject as one of many items in the stream of consciousness; but this is, as mentioned earlier, a mistake. The 'of'-talk leads us into a wrong picture. The way we talk about awareness (there is a subject and then there is something it is aware of) invites the mistaken thought that all cases of conscious awareness exhibit the structure of basic intentionality. There is an object of awareness in the case of primitive awareness too, but that 'object' is not given as any content in the stream of consciousness. The analogous observation applies to awareness of basic intentionality. Basic intentionality is the object of that conscious awareness but basic intentionality does not occur among other objects in what is present to the subject in that kind of awareness. One might have thought that, trivially, any object of any episode of conscious awareness must be an item within the totality of what is given to the subject (that the object at issue must occur as one object among others in the stream of consciousness). To think so is to believe that all kinds of conscious awareness are to be understood as episodes of phenomenal consciousness or, in other words, as cases of experiencing. It might be hard to agree that this is not so. Only by careful reflection upon primitive awareness and upon awareness of basic intentionality, one may come to see that this apparently unproblematic assumption should be abandoned.

The fundamental mistake of the representationalist can be found in the work of Brentano. It may now be described as follows: The idea that all cases of awareness of something are cases where an experience represents something quite clearly presupposes that conscious awareness is always a matter of being somehow phenomenally

presented with something. It quite clearly incorporates the intuition that all cases of conscious awareness have the structure of basic intentionality. But if the ideas here presented about primitive awareness and about awareness of basic intentionality are right, then the fact that representationalism incorporates this mistaken assumption explains why the representationalist does not have the conceptual resources to successfully tackle the problem recently rediscovered by self-representionalists: the problem about the way in which experiencing necessarily involves conscious awareness of experiencing and about the way in which it necessarily involves some kind of self-awareness. Since these are the fundamental problems about the nature of consciousness this remark amounts to the following diagnosis: Representationalist theories about the nature of consciousness are bound to fail.

18.11 Subjective Character: A Misleading Term

According to a widespread terminology that is rarely if ever put into question, qualitative character as well as the so-called subjective character are properties of experiences. A majority of philosophers presuppose furthermore and often without explicitly mentioning this presupposition that experiences are brain events. Presupposing the view here proposed we can now see in what way this terminology (the experience property framework) in particular when combined with the ontological assumption just mentioned leads astray. To see this, let us have a further look at each definition of subjective character proposed in this chapter.

According to the first definition, an experience has subjective character if and only if it is an event which consists in the fact that something specific is phenomenally given to the relevant subject. For an event to have subjective character is to fall into a specific metaphysical category, namely into the category of experiences, into the category of events involving an experiencing subject which has, in that event, changing experiential properties (something is phenomenally given to it). Subjective character in that sense raises deep problems about the nature of experiencing subjects and the nature of phenomenal presence. Introducing subjective character as a property of brain processes hides this fundamental insight. It is then easy to think that we are just searching for a certain special property and to overlook that we are addressing the deep and difficult question about the nature of certain kinds of individuals, experiencing subjects. Obviously, no philosophical account of subjective character in this first sense can be given without addressing these questions about the nature of these special individuals. The experience property framework makes it easy to overlook that platitude. It makes it easy to believe, as many contemporary philosophers apparently do, that we can seriously try to develop a philosophical account of consciousness, leaving the experiencing subject out of the picture. Talking of basic intentionality in terms of subjective character understood as a property of experiences (brain events) makes it appear—on the surface—as if we were indeed able to talk about an interesting and fundamental feature of consciousness here, remaining completely silent about the one who has the experience,

the experiencing subject. In reality, this is an illusion. If we use the term subjective character in the sense of definition 1 then we are in fact talking about the experiencing subject (the 'property of experiences' at issue is defined by reference to it), but we do so in a hidden and implicit manner as if we were talking in a secret code in order to hide our real communicative intentions. This way to proceed is obviously undesirable. Progress in philosophy is more likely if the participants are as open and clear as possible about what they actually mean.

A related point concerns the choice of basic terms. A developed precise philosophical theory about consciousness would have to include a proposal with respect to the language in which it is couched. This proposal would have to make it clear what terms are used as basic and what terms are introduced by definition. It would be supplemented by a careful explanation (using examples and theoretical assumptions) of its basic terms. Obviously, it is very unlikely that we will ever reach, in philosophy, the state where the ideal of such a precisely formulated theory will be attained. But still, this ideal should guide the way we proceed. If we use 'subjective character' in the sense of definition 1, then we should make it explicit that the term is a defined term and that its definition *requires* talking of experiencing subjects and phenomenal presence. It then would be obvious that these more fundamental terms need to be in the centre of the inquiry and that it is a mistake to try to remain silent about what they refer to. A different way to proceed is to introduce subjective character (in the sense of definition 1) as a fundamental term and to mention what occurs in definition 1 as the definiens only in passing by way of intuitive comments about what it is supposed to capture. But this way to proceed is not recommendable. It distracts attention from what should be in the centre of the philosophical enterprise.

In the sense of definition 2, the subjective character of an experience is for it to be such that the subject concerned is primitively aware of undergoing the experience. Here again, the property of the experience is introduced by reference to the experiencing being at issue and what has been said in the preceding paragraph reapplies. The fact that we cannot understand what subjective character in that second sense amounts to without reference to the experiencing subject should better not be hidden by the chosen terminology; it should rather be made explicit so that it is clear to everybody involved in the discussion that the old question about the bearer of experiential properties has reoccurred.

Finally, to talk of subjective character as a property of experiences having the sense explicated by definition 3 in mind, contains a double implicit reference to the experiencing subject and so hides reference to the subject twice: An experience has subjective character in this third sense if *the subject concerned* is aware of the metaphysical structure of the experience in which it is involved (first reference to the experiencing subject) and in being so aware of basic intentionality, the subject is aware of *itself* (second reference to the experiencing subject concerned).

When the experience property framework is combined with the presupposition that experiences are brain events then it is likely to mislead in yet another way. The risk is that those reflecting upon philosophical issues surrounding the topic 'subjective character' think of awareness of one's own experiences as a case of looking inside and discovering inner events with specific properties. The problem is that the

framework invites the mistaken perceptual metaphor of 'inner awareness'. It should be clear that neither primitive awareness of one's own properties nor pre-reflexive nonconceptual awareness of basic intentionality is the result of some kind of 'looking inwards'. And it should be clear (as pointed out earlier) that none of these two cases of awareness involve that some item is phenomenally present to the subject or given—so to speak—'as an object' to the experiencer. Being primitively aware of one's own experiencing is not a matter of being presented with experiences as a result of looking inside. Nor is awareness of basic intentionality a matter of discovering, inside oneself, things with a certain structure. Talking of properties of brain events in this context invites, however, this mistaken picture. In other words, the framework makes it easy to overlook that these two kinds of awareness do not have the structure of basic intentionality, that they are no cases of experiencing something.

Furthermore, the framework is likely to lead into even deeper confusion. Suppose that we take subjective character (a) to consist in some kind of 'inner' awareness of the experiencer of his or her own experience (it consists in the fact that the experiencer encounters that experience by 'looking inside') and that (b) we take subjective character to be a property of the event discovered inside via that 'inner look'. The second part (b) is likely to lead us to the idea that the experiencer who is aware of his or her experience via 'inner awareness' is then aware of an event with an interesting property: subjective character. It then looks as if subjective character is one of those properties the event discovered inside appears to have, when one is aware of it via 'the inner look' at issue. It then looks as if the experiencer, in being 'inwardly aware' of his or her own experience, discovers an event which appears to have subjective character. All this is already completely misguided but let us stay within this bad perceptual metaphor of inner awareness and see what happens when it is taken seriously. The experiencer then discovers an event which appears to have subjective character but that apparent property consists, according to (a), in the fact that the experiencer managed singling out that event via inner awareness. It follows that the inner event 'appears to the experiencer' in a particular way in virtue of the fact that he or she is 'looking at it' via inner awareness. This thought is confusing or rather confused and we should not try to make sense of it. The whole picture of inner awareness as awareness of inner processes should, of course, be abandoned. This has been pointed out repeatedly by a number of philosophers.[10] The perceptual metaphor of inner awareness is, however, still at work, or so it seems obvious to me, in contemporary thinking about consciousness and it is, once again, an obstacle in the ongoing debate about the so-called subjective character: Talking of subjective character in terms of properties of inner processes is likely to reactivate that bad metaphor in the thinking of many.

According to what has been said in this chapter, the term 'subjective character' is used to attract the attention of contemporary philosophers to fundamental issues about consciousness which have long been neglected within the analytical tradition.

[10] Compare, e.g. Sidney Shoemaker (1994) and the discussion of his views in Cynthia MacDonald (1999). In M. Nida-Rümelin (2007) I argue that the perceptual model is responsible for a number of fallacies in the philosophical debate about representationalism and qualia.

The term is, however, used in a systematically ambiguous manner. Furthermore, it is part of a conceptual framework (the experience property account) which is in various ways seriously misleading. For these reasons, although the themes raised by the discussion about subjective character deserve close attention and careful examination, the term itself should rather be abandoned and replaced by a series of different concepts that remain to be developed.

Acknowledgments I presented some of the material of this chapter in a workshop in Geneva on perception organized by Pascal Engel in November 2009 and in a workshop with Terrence Horgan in Fribourg organized by Emmanuel Baierlé in July 2010. I am indebted to the discussants on these occasions, in particular, to Kevin Mulligan and Terrence Horgan. In getting clear about the so-called subjective character, the discussions in my seminars in 2010 and 2011 were of great help; I would like to thank the participants for the fruitful exchange. Personal discussion and written exchange helped me a lot in the development of the ideas here presented; in particular, I would like to thank Emmanuel Baierlé, Julien Bugnon, Gianfranco Soldati, Max Drömmer and Daniel Stoljar. A lot of what is presented here developed in the discussions with Fabrice Theler about representationalism in the context of his work for the project 'First person access, phenomenal reflection and phenomenal concepts' supported by the Swiss Science Foundation (PDFMP1 132455). I would like to thank the Swiss Science Foundation for its support of our research.

Appendix: Citations with Comments

> 1. Now there are two features of conscious sensory states that require theoretical elucidation: 'qualitative character' and 'subjectivity'… In the case at hand, seeing a ripe tomato, there is both a distinctive qualitative character to be reckoned with and also the fact that the state is conscious — 'for the subject', in a way that unconscious states are not. (Levine 2006, Sect. 2, § 2)

Comment We can read this passage as follows: The difference between conscious and unconscious sensory experiences lies in the fact that the former but not the latter are in an important sense 'for the subject'. This may be taken to mean that being in the state makes a difference for the subject (primitive awareness); or it may be taken to mean that in a conscious state, contrary to an unconscious state, there is a subject to whom something is phenomenally present (basic intentionality). Both interpretations appear plausible and both interpretations are well compatible with the claim that being 'for the subject' marks the difference between conscious and unconscious sensory experiences.

> 2. I consciously see a ripe tomato on the kitchen counter. Clearly the primary object of my conscious state is what I'm seeing, the scene on the counter. But it also seems to be the case that my very awareness of the tomato on the counter includes within it somehow an apprehension that I am seeing what I'm seeing. (Levine 2006, Sect. 2, § 8)

Comment 'Apprehension that I'm seeing what I'm seeing' in the sense Levine has in mind in this passage is not the result of reflection upon one's experience. It is the awareness of seeing something which the subject has, necessarily, in seeing something. 'Apprehension' in this passage, I suggest, is best understood in the sense of primitive awareness.

> 3. To a first approximation, the experience's bluish qualitative character is what makes it the experience it is, but its for-me-ness is what makes it an experience at all. A better, if initially less clear, approximation is this: my experience is the experience it is because it is bluish-for-me, and is an experience at all because it is somehow-for-me (or qualitatively-for-me). Thus qualitative character is what varies among conscious experiences, while subjective character is what is common to them. (U. Kriegel, forthcoming in Lui and Perry (eds), Sect. 2, § 2)

Comment 'Bluish character' here is used as if it referred to a qualitative property of the experience (the relevant event). This technical expression, in order to be given a sense, must be translated into a talk about experiencing subjects and what they are presented with or what they are aware of. The only way, or so I claim, to understand what it is for an experience to be bluish (or to have a bluish character) is to say this: An experience has that property iff the subject involved in the event is phenomenally presented with blue. (It is common, following Levine (2001), to reserve the term 'bluish' for the property of an experience corresponding to 'blue'. Where 'blue' is the property things appear to have in the relevant colour perceptions and 'bluish' is the alleged property of experiences shared by those experiences where the colour blue is phenomenally present. This sense of bluish must not be confused with its normal sense where it characterizes the commonality between, e.g. violet and turquoise.) Since 'for-me-ness' is introduced, just like bluishness, as a property of the experience, one might try using the same schema for a translation into a less technical and more accessible language: An experience has that property iff in having the experience the subject is phenomenally presented with 'being for me'. What could it be, however, to be phenomenally presented with 'being for me'? In the case of 'bluish', we have an understanding of the corresponding property (blue) which can be given in an experience (normally as a property something appears to have) and we have a clear understanding of what it is to be phenomenally presented with that property. Therefore, the technical talk of bluishness can be eliminated and replaced by an unproblematic terminology. It is not possible to proceed in the same why for 'for-me-ness'. Contrary to the colour case, for-me-ness is not a candidate for a property something appears to have in an experience; and there are no candidates for objects either that might appear to have for-me-ness; we are not presented with an experience which appears to have for-me-ness (Sects. 18.10 and 18.11 of the present chapter).

In the above citation, an experience's having subjective character is said to consist in the fact that having the experience is somehow for the subject. It is plausible and natural to understand this as the claim that having the experience makes a difference for the subject and so it is natural to interpret 'for-me-ness' in this passage in the sense of primitive awareness.

> 4. According to Levine and me, the deeply mystifying feature of phenomenal consciousness is that when I have a conscious experience, the experience does not occur only *in me*, but also *for me*. There is some sort of direct presence, a subjective significance, of the experience to the subject. (U. Kriegel, forthcoming in Lui and Perry (eds.), Sect. 2, § 3)

Comment I propose to interpret this talk of being 'for me' as primitive awareness.

> 5. …for a conscious experience to be not only *in* me, but also *for* me, I would have to be *aware* of it. (U. Kriegel, forthcoming in Lui and Perry (eds.), Sect. 2, § 5)

Comment In my view, the only way to interpret 'awareness' here is to read it in the sense of primitive awareness. But then it is a mistake to go on and say that, therefore, the experience is itself represented in some state of the subject (as Kriegel does in the text which follows the cited passage). Primitive awareness of an experience is not a matter of representing the experience (compare Sect. 18.9 of the present chapter).

> 6. Not only *is* the experience bluish, but I am also *aware* of its being bluish. Its *being* bluish constitutes its qualitative character, while my *awareness* of it constitutes its subjective character. (Kriegel 2005, p. 27)

Comment The last occurrence of 'it' in this passage is puzzling. Does it refer to 'the experience's being bluish' (as the preceding sentence appears to suggest) or does it refer to the experience itself (as grammar appears to suggest)? But translating technical terminology in a more explicit language we can see that the difference does not matter. Being bluish (for an experience) means that the subject involved in the experience is phenomenally presented with blue. Being aware of the bluishness of an experience then must mean: being aware of being phenomenally presented with blue. And if the experience at issue consists in the subject's being phenomenally presented with blue, then being aware of the experience means, once again, being aware of being phenomenally presented with blue. I conclude that subjective character, as the term is used in this passage, should be interpreted in the sense of primitive awareness.

The language chosen to express primitive awareness is, however, misleading. To say that I am aware of my experience's being bluish clearly invites the perceptual metaphor discussed earlier (compare Sect. 18.11). One is invited to think of the experience as something phenomenally given in the experience (perhaps upon 'looking inside') as having the property of being bluish—which is, of course, false. No experience is phenomenally given; and no experience is phenomenally given as having certain properties. Being primitively aware of being presented with blue does not involve anything like that.

> 7. When a phenomenally conscious state represents something, it makes the subject aware of what it represents. To say that your perceptual experience represents both the page and itself is therefore to say that it makes you aware not only of the page, but also of your experience of the page. Let us call your awareness of the page *outer awareness* and your awareness of the experience *inner awareness*. (Horgan and Kriegel 2007, Sect. 2.1, § 2)

Comment This passage can be read as a short and apparently obvious argument for a representationalist account of the way in which we are necessarily aware of undergoing an experience in undergoing the experience. It presupposes that awareness of something is, in every case, a matter of representation. We should however, or so I claim, abandon this representationalist dogma (compare Sect. 18.9 of this chapter).

The term ‘inner awareness’ introduced here is likely to evoke the bad perceptual metaphor mentioned in Sect. 18.11.

> 8. In the ordinary go of things, one does not become aware of one’s ongoing experience through an extra mental step that results in the formation of a numerically distinct state of awareness. Rather, the awareness of the experience is a component of the experience itself, not a further mental event or state. (Horgan and Kriegel 2007, Sect. 2.1., § 3)

Comment This passage is an example of a widespread habit to discuss phenomenological issues about the correct description of experience in terms of questions about counting experiences and in terms of parts or components of experiences. But how should we count experiences and what depends on how many experiences are involved in a given case? Do we have any clear understanding of what it is for experiences to be numerically identical or distinct? Does it make sense to ask, for instance, how many experiences you had in the last five minutes? Talking in this way presupposes, in my opinion, a problematic reification of experiences. Furthermore, talking of parts or components of experiences in this context is just a metaphor. Both ways of talking can serve their purpose only because we have, implicitly, some translation into a more accessible language in mind. An adequate translation, I propose, can be formulated like this: When a subject undergoes an experience, then, in having the experience, the subject is necessarily aware of having it. Using the notion of phenomenal presence (and presupposing that every experience is a matter of being phenomenally presented with something), we get an even simpler claim: In being phenomenally presented with something the subject is necessarily aware of being phenomenally presented with that something. This simpler affirmation is more easily accessible for intuitive testing on the basis of phenomenological reflection. Nothing appears to be lost if we abandon the technicalities present in the above citation in favour of the proposed simpler language and a lot is gained: The intuitive content becomes clearly visible.

> 9. Consider your perceptual experience of this page. It makes you aware primarily of the page, not of itself. This is because your attention is absorbed with the page. Yet, we maintain, you are also aware, though much more dimly, of having that very experience. (Horgan and Kriegel 2007, 2.1., § 7)

> 10. Thus it seems all but incoherent to suppose that one could have a phenomenal experience which was greenish, but of which one was aware as reddish. For, what it is like *for the subject* to have the experience is determined by the way the subject is aware of her experience. If the subject is aware of the experience as reddish, then what the experience is like *for the subject* is reddish. (In the ordinary case, the subject is focally aware of an external object as red, via an experience deploying a reddish mode of representation of that red object; the subject thereby is peripherally aware of the experience itself as reddish, since the reddish experience represents both the red external object and itself.) (Horgan and Kriegel 2007, 2.1, § 11)

Comment The above two citations contain the idea that the kind of awareness at issue is peripheral, an idea criticized in Sect. 18.7 of the present chapter. Neither basic intentionality, nor primitive awareness, nor awareness of basic intentionality allow for any distinction between being more or less peripheral.

To be aware of an experience as reddish can be translated in a simpler and less problematic language: it is to be aware of being phenomenally presented with red. To say then that one cannot have a greenish experience and be aware of it as reddish means, quite simply, this: You cannot be phenomenally presented with green and, in being so presented with green, be aware of being phenomenally presented with red.

> 11. On my view, however, there is more to be said about phenomenal character —there is more structure to it than is typically recognized. In particular, I distinguish two components of the 'bluish way it is like for me' to have the experience: the bluish component, which I call *qualitative character*, and the for-me component, which I call *subjective character*. To make a conceptual separation between qualitative and subjective character is not to imply that they can occur apart from one another.
> My view is that there are many determinate phenomenal characters —bluish-for-me-ness, greenish-for-me-ness, bitterish-for-me-ness, trumpet-for-me-ness, etc.— and the determinable of all of them is for-me-ness as such. We grasp what subjective character is by fixing on what is common to all phenomenally conscious states, and grasp what qualitative character is by fixing on what varies among them. (U. Kriegel, forthcoming in *Philosophical Studies*, first page)

Comment For an experience to be 'bluish' means that the subject involved is phenomenally presented with blue. For a 'bluish' experience to have 'for-me-ness' is for the subject to be aware of being presented with blue. To pack both parts in one alleged property of an experience is confusing. It masks the fact that its first part and its second part must be translated into the language of ordinary thought in quite different ways. Talking of 'bluish-for-me-ness' invites the idea that there is one single property such that for an experience to have bluish-for-me-ness is for the subject to be phenomenally presented with that property. But there is no such property. The subject is only presented with blue and there is no phenomenally presented property corresponding to the 'bluish-for-me-ness' as a whole or to the 'for-me-ness' part alone.

Using the terminology introduced in the present chapter, we can say what all phenomenally conscious states have in common: (a) they exhibit basic intentionality (there is a subject to whom something is phenomenally given), (b) the subject is primitively aware of being in that state and (c) the subject is, in having the experience, aware of its basic intentionality. According to the view presented, (a), (b) and (c) are all necessarily fulfilled in each phenomenally conscious state. Furthermore, each of them necessarily implies the two others. Therefore, we can use each of them as the mark of the phenomenal. So each of the three interpretations of subjective character proposed in the present chapter is perfectly compatible with Kriegel's claim that subjective character is what all phenomenally conscious states have in common.

In my interpretation, qualitative character is characterized by what is phenomenally present to the experiencing subject. This interpretation too complies with Kriegel's claim that we grasp qualitative character when we focus on what distinguishes different phenomenally conscious states: Two phenomenally conscious states differ in virtue of differences in what is phenomenally present to the experiencing subject.

> 12. In demselben psychischen Phänomen, in welchem der Ton vorgestellt wird, erfassen wir zugleich das psychische Phänomen selbst, und zwar nach seiner doppelten Eigentümlichkeit, insofern es als Inhalt den Ton in sich hat, und insofern es zugleich sich selbst als Inhalt gegenwärtig ist.
> Wir können *den Ton das primäre*, das Hören selbst das *sekundäre Objekt* des Hörens nennen. (Brentano 1874, 179/180)
> English translation: In the same mental phenomenon in which the sound is present to our minds we simultaneously apprehend the mental phenomenon itself. What is more, we apprehend it in accordance with its dual nature insofar as it has the sound as content within it, and insofar as it has itself as content at the same time.
> We can say that the sound is the *primary object* of the *act* of hearing, and that the act of hearing itself is the *secondary object*. (Cited from the translation by L. L. McAlister. *Psychology from an Empirical Standpoint*, London: Routledge, 1973)

Comment Like contemporary authors, Brentano does not explicitly talk of the subject to whom something is presented; he says that the experience (‘das psychische Phänomen’/‘the mental phenomenon’) is presented to itself as a content and that it is as an object (a secondary object) of the experience. We may say then, or so it appears plausible to me, that Brentano attributes basic intentionality to the subject’s relevant kind of awareness of hearing a tone (to the kind of awareness included in the hearing itself). He appears to think that the relation involved here of being presented (‘vorgestellt’ or ‘gegenwärtig als Inhalt’) is not at all fundamentally different in the case of the awareness of the tone while hearing the tone and in the case of the awareness of hearing the tone while hearing the tone. We thus might translate him as saying that both, the tone and the hearing of the tone, are phenomenally present to the subject in hearing the tone. So Brentano appears to commit the representationalist mistake described in Sect. 18.9.

The representationalist mistake, as I have been trying to describe it in Sect. 18.9 of the present chapter, might be summarized saying something like this: The representationalist mistakenly assumes that when a subject is aware of its experience in having the experience, the experience itself is a further object of this awareness. An alternative to the representationalist view is to say that this particular kind of awareness, although it is ‘of something’ in a certain sense, nonetheless does not have an object. Zahavi (2006) ascribes this object-less view to Husserl, Sartre and Heidegger and he describes it as follows:

> Thus, not only does it [the Husserlian account] reject the view that a mental state becomes conscious by being taken as an object by a higher-order state, it also rejects the view espoused by Brentano according to which a mental state becomes conscious by taking itself as an object. Brentano and Husserl both share the view that self-consciousness (or to use Brentano’s terminology ‘inner consciousness’) differs from ordinary object-consciousness. The issue of controversy is over whether selfconsciousness is (i) merely an extraordinary object-consciousness or (ii) not an object consciousness at all. In contrast to Brentano, Husserl thinks the latter, more radical, move is required. (Zahavi 2006, p. 5)

According to the view proposed in the present chapter, neither primitive awareness nor awareness of basic intentionality has the structure of basic intentionality. If this is correct, then primitive awareness and awareness of basic intentionality are to be

seen in a way similar to Husserl's view about pre-reflexive consciousness. Here is another passage in which Zahavi describes the relevant aspect of Husserl's view:

> This might come as a slight surprise to those who thought that one of the central doctrines in phenomenology is the doctrine of intentionality —i.e., the claim that all consciousness is intentional, that all consciousness is object consciousness— but Husserl (as well as later phenomenologists) would explicitly deny that pre-reflective self-consciousness involves a subject-object relation. In his view, when one is pre-reflectively conscious of one's own experiences, one is not aware of them as objects. My pre-reflective access to my own mental life in first-personal experience is immediate, non-observational and non-objectifying. It is non-objectifying in the sense that I do not occupy the position or perspective of a spectator or in(tro)spector on it. (Zahavi 2006, p. 6)

References

Boghossian P, Velleman, D (1989/1997) Colour as a secondary quality. Mind 98:81–103. Reprinted in: Byrne, A, Hilbert, D (eds) Readings on color, Volume I: the philosophy of color. The MIT Press, Cambridge

Brentano F (1874) Psychologie von empirische Standpunkt, Vol. 1, book 2.

Cramer K (1974) Erlebnis. Thesen zu Hegels Theorie des Selbstbewußtseins mit Rücksicht auf die Aporien eines Grundbegriffs nachhegelscher Philosophie. In: Gadamer HG (eds) Stuttgarter Hegel-Tage 1970. Hegel-Studien, Bonn, pp. 537–603 (Beiheft 11)

Crane T (2002) Introspection, intentionality, and the transparency of experience. Philos Top 28(2):49–67

Dorsch F (2011) Transparency and imagining seeing. Philos Explor 13(3):173–200

Frank M (2011) Ansichten der Subjektivität. Suhrkamp Taschenbuch Wissenschaft

Gallagher S, Zahavi, D (2004/2010) Phenomenological approaches to self-consciousness. Stanford Encyclopedia of Philosophy

Gurwitsch A (1977) Die mitmenschlichen Begegnungen in der Milieu-Welt, Volume 16 of "Psychologische und Phänomenologische Forschungen". Walter de Gruyter. English edition: Gurwitsch, A, Human Encounters in the Social World (trans: Kersten, F). Duquesne University Press, Pittsburgh

Henrich D (1970) Selbstbewußtsein, kritische Einleitung in eine Theorie. In: Bubner R, Cramer K, Wiehl R (eds) Hermeneutik und Dialektik. Mohr, Tübingen, pp 257–284

Horgan T, Kriegel U (2007) Phenomenal epistemology: what is consciousness that we may know it so well? Philos Issues 17(1):123–144

Husserl E (1900/1901/1984) Logische Untersuchungen II, Husserliana XIX/1-2. Den Haag, Martinus Nijhoff. English edition: Husserl, E (2001) Logical Investigations I-II (trans: Findlay, JN). Routledge, London

Kind A (2003) What's so transparent about transparency? Philos Stud 115(3):125–244

Kind A (2010) Transparency and representationalist theories of mind. Philos Compass 5:902–913

Kriegel U (2005) Naturalizing subjective character. Philos Phenomenol Res 71(1):23–57

Kriegel U (2009) Subjective consciousness: a self-representational theory. Oxford University Press, Oxford

Kriegel U (forthcoming) Précis of "subjective consciousness: a self-representational theory". Philosophical Studies

Kriegel U (forthcoming) Self-representationalism and the explanatory gap. In: Liu J, Perry J (eds) Consciousness and the self. Cambridge University Press, Cambridge

Levine J (2001) Purple haze. Oxford University Press, Oxford

Levine J (2006) Conscious awareness and (self-) representation. In: Williford K, Kriegel U (eds) Self-representational approaches to consciousness. The MIT Press, Cambridge

MacDonald C (1999) Shoemaker on self-knowledge and inner sense. Philos Phenomenol Res 59(3):711–738
Martin MGF (2002) The transparency of experience. Mind Lang 4(4):376–425
Moore GE (1914) The status of sense-data. Proc Aristotelian Soc, New Series 14:355–380
Mulligan K (2004) Brentano on the mind. In: Jacquette D (ed) Cambridge companion to Brentano. Cambridge University Press, Cambridge, pp. 66–97
Mulligan K, Smith, B (1985) Franz Brentano on the ontology of mind. Philos Phenomenol Res 45:627–44
Nida-Rümelin M (2007) Transparency of experience and the perceptual model of phenomenal awareness. Philos Perspect 21(1):429–455
Nida-Rümelin M (2011) Phenomenal presence and perceptual awareness: a subjectivist account of perceptual openness to the world. Philos Issues 21(1):352–383
Pothast U (1971) Über einige Fragen der Selbstbeziehung. Vittorio Klostermann, Frankfurt a. M.
Sartre JP (1936) La transcendance de l'ego. Vrin, Paris
Shoemaker S (1994) Self-knowledge and inner sense. Philos Phenomenol Res LIV:249–314
Siewert C (2004) Is experience transparent? Philos Stud 117(1–2):15–41
Stoljar D (2004) The argument from diaphonousness. In: Escurdia M, Stainton RJ, Viger CD (eds) Language, mind and world: special issue of the Canadian journal of philosophy. University of Alberta Press, Alberta
Tye M (2002) Representationalism and the transparency of experience. Noûs 36(1):137–151
Zahavi D (2006) Two takes on a one-level-account of consciousness. Psyche 12(2):1–9

Part III
Philosophy of Mind and Philosophy of Language

Chapter 19
Causal Equivalence as a Basis for the Specification of Neural Correlates

Uwe Meixner

Abstract This paper defines causal equivalence and uses it to explicate the notion of neural correlate. It is seen that many issues of the mind–body problem can be fruitfully discussed in terms of causal equivalence, or in relation to it. The paper ends by proposing a dilemma that cannot be rationally resolved.

Keyword Causal equivalence · Neural correlate · Psychophysical dualism · Causal closure · Causal overdetermination

19.1 Two Concepts of Causal Equivalence

Let E and E' be events:

(a) E' is (simpliciter) *causally equivalent* to E if, and only if, (i) everything caused by E is also caused by E', and *vice versa* (substitute the statement that is obtained by interchanging "E" and "E'" in the previous statement), and (ii) everything that causes E also causes E', and vice versa.
In other words: E' is causally equivalent to E if, and only if, E' has the same effects and the same causes as E.
(b) E' is an *event-causally equivalent* to E if, and only if, (i) every event caused by E is also caused by E', and vice versa, and (ii) every event that causes E also causes E', and vice versa.
In other words: E' is an event-causally equivalent to E if, and only if, E' has the same event effects as E and the same event causes as E.

Notes:

(i) *Causal equivalence* entails *event-causal equivalence*, but *not* vice versa—unless, of course, *causation* coincides with *event causation*; this is believed by many, but is not proven. In what follows, I shall stick to the simple concept of causal equivalence, without argumentatively deciding whether it coincides with event-causal equivalence or not (being a believer in agent causation, I do believe that the two concepts of causal equivalence do not coincide).

U. Meixner (✉)
University of Augsburg, Augsburg, Germany
e-mail: uwe.meixner@phil.uni-augsburg.de

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_19,

(ii) Every event is causally equivalent to itself; and if an event *E′* is causally equivalent to an event *E,* then *E* is also causally equivalent to *E′;* and if an event *E″* is causally equivalent to an event *E′,* which is in turn causally equivalent to an event *E,* then *E″* is also causally equivalent to *E*. Thus, the relation of causal equivalence is reflexive, symmetrical, and transitive among events; in other words: it is an equivalence relation among events.

19.2 Two Theses of Causal Equivalence Regarding Mental Events

(I): For every actual mental event *E:* there is some wholly physical actual event *E′* such that *E′* is causally equivalent to *E*.
(II): For every actual mental event *E:* there is some (at least) partly physical actual event *E′* such that *E′* is causally equivalent to *E*.

Note:

Being wholly physical entails *being* (*at least*) *partly physical*, but not vice versa. Hence, thesis (I) entails thesis (II), but not vice versa.

19.3 Two Theses of Psychophysical Dualism

(1) Some actual mental event is wholly nonphysical.
(2) Some actual mental event is not wholly physical.

Notes:

(i) Thesis (1) entails thesis (2), but not vice versa.
(ii) *Wholly nonphysical* is logically equivalent to *not partly physical; not wholly physical* is logically equivalent to *partly nonphysical*.

19.4 Two Theses of Causal Non-Equivalence Regarding Mental Events

Non-(I): For some actual mental event *E:* there is *no* wholly physical actual event *E′* such that *E′* is causally equivalent to *E*.
Non-(II): For some actual mental event *E:* there is *no* partly physical actual event *E′* such that *E′* is causally equivalent to *E*.

Notes:

(i) Thesis *non*-(II) entails thesis *non*-(I), but not vice versa.
(ii) Thesis *non*-(I) entails the dualistic thesis (2) (that is, *some actual mental event is not wholly physical*).
(iii) Thesis *non*-(II) entails the dualistic thesis (1) (that is, *some actual mental event is not partly physical*, i.e., *is wholly nonphysical*).

19.5 Two Research Programs for Cognitive Neuroscience

(A) Seek to *corroborate* thesis (I) (and hence also thesis (II)) to the point of *definitely establishing* thesis (I) (and hence also thesis (II))—that is, in the sense a scientific hypothesis can be said to be "definitely established." Or, in default of reaching this goal, seek to *corroborate* thesis *non*-(I) to the point of definitely establishing it; if this other effort is crowned with success, the dualistic thesis (2) will have been established.
(B) Seek to *corroborate* thesis (II) to the point of definitely establishing it. Or, in default of reaching this goal, seek to *corroborate* thesis *non*-(II) to the point of definitely establishing it; if this other effort is crowned with success, the dualistic thesis (1) will have been established (and also thesis *non*-(I)).

Notes:

(i) Obviously, each of these two research programs is very significant for the advancement of human knowledge.
(ii) Even if thesis (I) was definitely established, this does not mean that dualism stands refuted, not even in the stronger version represented by thesis (1). The reason for this is the following: even if, for a given actual mental event *E,* there is a wholly physical actual event *E′* that is causally equivalent to *E* (as must be the case according to thesis (I)), it does not follow that *E* is at least partly physical; *E* may still be wholly nonphysical.
(iii) Matters would be different if one had a sufficient reason to assume that *causally equivalent actual events are identical*. But, whereas it is very plausible to assume that causally equivalent *wholly physical* actual events are identical, it amounts to begging the question against dualism if one assumes that an actual mental event *E* and an actual wholly physical *E′* are identical if they are causally equivalent. On the contrary, it seems that there are noncausal, "inner" properties—properties that *E* has (qua mental event), but *E′* has not (qua wholly physical event)—which distinguish *E* from *E′* even if *E′* is causally equivalent to *E*.
(iv) Since causally equivalent wholly physical actual events are—*very plausibly*—identical, there cannot be *more than one* wholly physical actual event that is causally equivalent to a given actual mental event. (Suppose that both *E′* and *E″* are wholly physical actual events and are *both* causally equivalent to the actual mental event *E*; hence, they are also causally equivalent to each other—

the causal equivalence of events being a symmetrical and transitive relation—and *therefore* E' and E'' are identical, according to the invoked principle of identity for wholly physical actual events.)

(v) It may also be true that all causally equivalent actual *mental* events are identical; but perhaps some intrinsically different actual mental events are, though nonidentical, nevertheless causally equivalent. I leave it as an open question which of these two conflicting hypotheses is true.

19.6 Neural Correlates as Causal Equivalents?

As a matter of fact, scientists appear to be working already on research program (A). Given an actual mental event E (a visual experience, say), they are trying to find a—in fact, *the*—wholly physical actual event E'—a brain event, to be specific—which is *correlated* with E: E' is called "the neural correlate of E." The methods used in this scientific enterprise are still very coarse but they may improve with time. What should interest us *here and now* is the following question: what, precisely, is the nature of the invoked *correlation* between E' and E? Which fact of correlation makes a wholly physical actual event *the neural correlate* of an actual mental event?

A wholly physical actual event E' is selected as *the neural correlate* of an actual mental event E on the basis of three criteria: first, *spatial location:* E' is in that region of the brain of *the person with* E where the physical data connected with E are ultimately processed; second, *temporal location:* E' is simultaneous—or at least approximately simultaneous—with E; third, *unique prominence:* E' stands out—against all other physical events that are also simultaneous with E and are also located in the relevant brain region—in a manner that links E' uniquely with E.

Now, these three criteria for selecting a wholly physical actual event E' as *the neural correlate* of an actual mental event E, they all point in the direction of the neural correlate of E being nothing else than the wholly physical actual event that is causally equivalent to E:

> The neural correlate of E = the wholly physical actual event that is causally equivalent to E.

In fact, the truth of this identity statement is guaranteed if E has *some* neural correlate *and* if the following principle is adopted as a (partial) analysis of the predicate "X is a neural correlate of E."

> Principle of Neural Correlation (PNC)
> For all actual mental events E:
> For all X: X is a neural correlate of E if, and only if, X is a wholly physical actual event that is causally equivalent to E.[1]

[1] The identity statement (preceding PNC) follows from PNC and the assumption that (some) X is a neural correlate of the actual mental event E as follows: Suppose, X is a neural correlate of the actual mental event E. Hence by PNC: X is a wholly physical actual event that is causally equiva-

PNC seems attractive, one reason for its attractiveness being its relative clearness. But it has some not entirely obvious consequences that do not fit some assumptions that are often made regarding actual mental events and their causal relations to wholly physical actual events. *Let E be an actual mental event*; *then we have:*

Firstly, according to PNC: No neural correlate of *E* causes *E*. Suppose *X* were a neural correlate of *E* and caused *E*; it follows by PNC that *X* is causally equivalent to *E,* which entails, according to the definition of causal equivalence (Sect. 19.1), that everything that causes *E* also causes *X*. But, according to supposition, *X* causes *E*. Therefore, *X* causes *X*—which is impossible. Thus, we have *Theorem 1*: *Neural correlates of actual mental events do not cause the mental events of which they are neural correlates*.

Secondly, according to PNC: If there is a neural correlate of *E,* then *E* causes some wholly physical event. Suppose *X* were a neural correlate of *E* and *E* caused no physical event; it follows by PNC that *X* is a wholly physical actual event that is causally equivalent to *E,* which entails, according to the definition of causal equivalence, that everything caused by *X* is also caused by *E*. But, according to supposition, *E* causes no wholly physical event. Therefore, *X* causes no wholly physical event—which is certainly false, seeing that *X* is a wholly physical actual event: *every* such event causes some wholly physical event. Thus, we have *Theorem 2*: *No actual mental event with a neural correlate is causally epiphenomenal with regard to wholly physical events*.

Thirdly, according to PNC: If *E* has a neural correlate and is itself a wholly nonphysical event, then this results in a case of genuine causal overdetermination. Suppose *X* is a neural correlate of *E* and *E* is a wholly nonphysical event; it follows by PNC that *X* is a wholly physical actual event that is causally equivalent to *E*. And therefore, since *X* causes some wholly physical event (see the previous paragraph), *E* causes that same event, too—a situation which certainly constitutes a case of genuine causal overdetermination, *seeing* that *X* is a wholly physical event and *E* a wholly nonphysical one. Thus, we have *Theorem 3*: *Wholly nonphysical actual mental events with neural correlates give rise to genuine causal overdetermination*.

Fourthly, perhaps there are no cases of genuine causal overdetermination although one can certainly not exclude its occurrence a priori. In any case, on the basis of Theorem 3 we have *Theorem 4*: *If there is no genuine causal overdetermination, then actual mental events with neural correlates are not wholly nonphysical* (but at least partly physical).

Fifthly, we have—and this is perhaps the most significant result—*Theorem 5: If every event that causes a wholly physical event is itself wholly physical, then every actual mental event with a neural correlate is identical to that correlate*.[2]

---

lent to *E*. Hence (Sect. 19.5, note (iv)) only *X* is a wholly physical actual event that is causally equivalent to *E*. Hence by PNC: only *X* is a neural correlate of *E*. Hence, *X* is the wholly physical actual event that is causally equivalent to *E*, and *X* is the neural correlate of *E*. And therefore, the neural correlate of *E* = the wholly physical actual event that is causally equivalent to *E*.

[2] The proof is easy: Suppose every event that causes a wholly physical event is itself wholly physical, and suppose *E* is an actual mental event with the neural correlate *X*. Hence by PNC: *X* is a wholly physical actual event that is causally equivalent to *E*. Since *every wholly physical actual*

Theorem 5 is the basis of one of the great unresolved philosophical dilemmas of our time: If it is true that every event that causes a wholly physical event is wholly physical, then it cannot also be true that some actual mental event with a neural correlate is not identical to that correlate—no matter how much this proposition of non-identity may seem to be true. If, on the other hand, it is true that some actual mental event with a neural correlate is not identical to that correlate, then it cannot also be true that every event that causes a wholly physical event is itself wholly physical—no matter how fervently one may wish this proposition of physical causal closure to be true from a monistic point of view. You have the philosophical choice; I submit, there is nothing that will rationally determine how you should decide. So take your pick.[3]

event causes some wholly physical event (see the deduction of the Theorem 2), *X* causes a wholly physical event *Y*. Since *X* is causally equivalent to *E*, not only *X* but also *E* causes the wholly physically event *Y*, and therefore, according to supposition, *E* is a wholly physical event.But *wholly physical actual events that are causally equivalent are identical* (note (iv), Sect. 19.5). Therefore, *E* is identical to its neural correlate *X*.

[3] More on causal representation on the basis of causal equivalence—its relata being causal representatives of each other—and a demonstration of the relevancy of these concepts for the mind–body problem can be found in my book *The Two Sides of Being. A Reassessment of Psycho-Physical Dualism* (Paderborn, Mentis 2004). There I defend, among other things, a position I call interactionist parallelism.

Chapter 20
Simulation Versus Theory-Theory: A Plea for an Epistemological Turn

Julien A. Deonna and Bence Nanay

Abstract Simulation, if used as a way of becoming aware of other people's mental states, is the joint exercise of imagination and attribution. If *A* simulates *B*, then (1) *A* attributes to *B* the mental state in which *A* finds herself at the end of a process in which (2) *A* has imagined being in *B*'s situation. Although necessary, imagination and attribution are not sufficient for simulation: the latter occurs only if (3) the imagination process grounds or justifies the attribution. Depending on the notion of justification we use to make sense of the idea that an episode of imagining *serves as a reason* for attributing a mental state, the shape of the debate and the options it offers look very different. Reconfiguring the discussion in this way, we claim, shifts the focus of the simulation versus theory-theory debate to a question located in epistemology.

Keywords Simulation · Theory-theory · Imagination · Attribution · Epistemology

How do we become aware of other people's mental states? One possibility is that we are equipped with a theory whose domain of application is constituted by other agents' mental states. On this view, becoming aware of someone else's mental state is a case of inferring from a token behaviour the mental state that has caused it by applying the relevant part of a psychological theory. Another possibility is that we have the capacity to simulate other people's mental states; that is, we are able to put ourselves in other peoples' shoes, and go in imagination through the mental states we would go through were we really in the other person's circumstances. The end result of such a process, namely the mental state in which the simulator finds herself, can now serve as a guide to what mental state the simulated person is in.

In recent years, it has been widely acknowledged that these two models, the 'theory-theory' model and the 'simulation' model of our mind-reading abilities, are

J. A. Deonna (✉)
University of Geneva, Geneva, Switzerland
e-mail: julien.deonna@unige.ch

B. Nanay
University of Cambridge, Cambridge, UK
e-mail: bn206@cam.ac.uk

University of Antwerp, Antwerp, Belgium

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_20,

not so much competing theories of the same phenomenon but rather different means at our disposal to make sense of others as psychological beings. But this does not mean that the simulation versus theory-theory debate has become obsolete. On the contrary: we argue that it should not and could not be properly examined without bringing in wider epistemological considerations. In other words, rather than providing any direct arguments in favour of or against any of these models, we would like to contribute to the debate by clarifying its framework.

We are very much aware of the fact that the contemporary debate about the attribution of mental states to others is not an either/or debate between simulationists and theory-theorists, but rather between these two views on one hand and a new contestant, the 'phenomenological' or 'enactive' view on the other (Hutto 2007, 2008; Zahavi 2008; Ratcliffe 2007; Gallagher 2005, 2007a, b). According to this third option, we become aware of other agent's mental states without relying either on imagination or simulation or the explicit application of a theory (maybe by 'directly perceiving' their mental states). We will not say much about this third way of thinking about the attribution of mental states and restrict our argument to various versions of the simulationist and the theory-theory view, because the question of justification play a much clearer role in these. It would be interesting to examine the epistemological commitments of the 'phenomenological'/'enactive' view, but we will not do that here.

After examining what we take to be two crucial concepts in the specification of what simulating consists in, imagination and attribution, we argue that the focus of the simulation versus theory-theory debate revolves around the following question: In what way can a subject *A*'s imagining herself in subject *B*'s situation constitute a reason for her attribution of a mental state to *B*? If the answer to this question is that it is the application of a psychological theory that licenses this step then we end up with an account of our mind-reading abilities that draws on both simulation and theory: we get a 'hybrid account'. If, on the other hand, the answer is that imagination can be a reason for attribution without recourse to any theory, then we end up with an account that we could call 'pure simulation'. The result of reconfiguring the debate in that way shifts the focus of the debate away from the topic of our mind-reading abilities to a question located in epistemology.

20.1 Simulation: The Joint Exercise of Imagination and Attribution

A standard way of characterizing simulation is the following: An agent *A* imagines herself in *B*'s circumstances, gets a grip on what she, *A*, would do (see, feel, think, etc.) and concludes that this is what he, *B*, would also do (see, feel, think, etc.) in these circumstances.[1] As Gregory Currie writes: 'I imagine myself to be in the other person's position, […] I simply note that I formed, in imagination, a certain belief, desire or decision, then attribute it to the other' (Currie 1995, pp. 144–145).

[1] *A* and *B* can be the same person.

What this approximate characterization hides is that we engage in simulation in very different ways and for very different reasons. Sometimes we simulate others because we believe it is the best way to know what they will do (see, feel, think, etc.), but at other times, we simulate someone else just in order to coordinate our movements with them. Sometimes we simulate with the aim of getting to know what we should be doing at some point in the future, which is what happens if we are planning to perform some action, but at yet other times, we simulate in order to confirm or specify further an attribution that is already in our possession. Some further cases of simulation might constitute educational or recreational projects.[2] These aims often combine.

Whatever the goal is, one attributes to someone else a mental state not only as a result of having imagined being in the situation she is in, but also *because* of it. *A* simulates *B* if and only if:

1. *A* imagines being in *B*'s situation
2. *A* attributes a mental state to *B*
3. The *reason* for (2) is (1): the *reason* for *A*'s attributing a mental state to *B* is *A*'s imagining herself in *B*'s situation

This chapter will focus on the manner in which imagination can constitute a reason for attribution.[3]

Before turning to these questions, however, we need to elaborate on our characterization of simulation by examining its two central components: imagination and attribution, keeping in mind that the characterization of these concepts has to be broad enough to cover the multifaceted phenomenon that has been called simulation. We shall then be in a position to test our account against the ways in which simulation is understood in the literature (Sect. 20.2). Finally, we will show how the way in which we interpret the epistemic notions of reason or justification at play in clause (3) above illuminates an important aspect of the debate between the simulationist and the theory-theorist (Sects. 20.3–20.5).

20.2 Imagination and Attribution

Imagining being in someone else's situation is, to use Kendall Walton's phrase, a case of imagining *de se*. He writes:

[2] We sometimes imagine ourselves in counterfactual situations for the sheer pleasure of it. In this type of cases, the imaginative project is not undertaken as form of mind reading: no mental state is attributed to ourselves or to anyone else. Although the term 'simulation' is often used to cover the latter cases too, here we reserve the term to refer to forms of *mind-reading* activities—for which it was introduced.

[3] Note that the problem we are dealing with here is not part of the global skeptical worry with regard to the existence of other minds. Our question is rather: given that others have mental states, what is the nature and the role of simulation in our becoming aware of them.

> 'Imagining de se' [is] a form of self-imagining characteristically described as imagining *doing* or *experiencing* something (or *being* a certain way), as opposed to imagining merely *that* one does or experiences something or possesses a certain property. (Walton 1990, p. 29. Original emphasis)

If I imagine being in your situation, then I do not simply imagine *that* I am in your situation, but I imagine experiencing what I would experience were I in your situation.

Imagining being in someone else's situation is often interpreted as a special case of imagining *de se,* namely, 'imagining from the inside'. An episode of 'imagining from the inside' can be broadly characterized as one in which I imagine myself in your situation from your point of view or perspective (see Walton 1990 on a detailed discussion of imagining from the inside).

Adopting someone else's perspective could mean at least two things. First, I can attempt to adopt your perspective in the sense that I transpose myself, in imagination, into a *physical* space from where you witness the situation.[4] Second, I can attempt to adopt your *psychological* make-up or dispositions in the circumstances.[5] Part of the reason why the notion of 'imagining from the inside' is so notoriously vague[6] may be that the notion of perspective can be interpreted in either of these two senses of perspective.

When we imagine ourselves in other people's situation, we may or may not constrain the process on these two dimensions of perspective. In other words, we may or may not imagine the other person 'from the inside'. In the spirit of accounting for as much as possible of what simulation covers, we shall remain as liberal as possible in our characterization of imagining oneself in someone else's situation. Sometimes this amounts to imagining the other person from the inside (in one or both senses of the term), but sometimes it does not. The concept of imagination relevant for simulation is *imagining oneself being in someone else's situation*.

Three observations are now in order. First, given the various goals we pursue when simulating, the phrase '*B*'s situation' has to be interpreted broadly. The situation in question might be actual, counterfactual, future conditional or even fictional. Second, on this picture, imagining being in someone else's situation is not necessarily a case of attempting to replicate the experiences of another person, that is, it is not a process that fails or succeeds depending on the actual degree of similarity between *A*'s imagined mental states and *B*'s actual mental states. In other words,

[4] This is the subcase of imagining having someone else's experiences that Wollheim calls 'central imagining'. Wollheim explicitly argues that there are other cases of imagining having someone else's experiences, namely, examples of 'acentral imagining'. Like Wollheim, we regard both cases as imagining having someone else's experiences (Wollheim 1987, p. 103 ff.; See also Wollheim 1974, especially pp. 179–80; Wollheim 1984, pp. 72–83). See also Gregory Currie on 'impersonal imagination' (Currie 1995, pp. 155–180), and Peter Goldie on 'empathy' (Goldie 2000, 2001).

[5] See, for example, Smith (1995, 1997) who insists that for imagining from the inside, there must be some (though not necessarily exhaustive) similarity between the experiences of the imaginer and the person imagined from the inside.

[6] Wollheim famously wrote that he "could not trust that phrase [the phrase 'imagining from the inside'] so abused in philosophy" (Wollheim 1974, p. 87).

imagining myself in your situation is really imagining myself in *what I take to be* your situation.[7]

Finally, although imagining can be taken to be an intentional action, we do not want to be committed to this narrow way of thinking of imagining in general and imagining oneself in the other person's situation in particular. Imagining can be unintended and even unconscious. Again, we would like to keep the imagination component of simulation as general as possible, therefore not restricting the concept of imagining to an intentional conscious mental action.

The concept of attribution is less complicated. The easiest way (and perhaps the most typical in the literature) in which to think of *A*'s attributing a mental state to *B* is as *A*'s representation of *B*'s mental state: *A* represents *B* having a certain mental state.[8]

We can now put together our respective characterizations of imagination and attribution. (a) I imagine myself being in your situation, (b) I attribute a mental state to you, but, crucially, (c) the reason for my attributing this mental state to you is that I imagine myself being in your situation. In other words, the reason for (b) is (a).

Before we examine further clause (c) and the ways in which this analysis may shed light on the debate between the simulationist and the theory-theorist, we should explore how well our account of simulation captures the various ways in which this concept is conceived of in the literature.

Simulation theories come in many varieties. Adam Morton identified three key dimensions around which a rough taxonomy of different conceptions of simulation could be devised (Morton 2003, pp. 122–135). Different theories tend to emphasize and combine these different key dimensions in different ways. What is important for our purposes is to show that our account of simulation accommodates any combination of these dimensions.

The first dimension along which a theory of simulation might vary concerns the level of the cognitive processes that it involves. Simulation might be thought of as operating at a subpersonal level, or, alternatively, at the personal level (Morton 2003, p. 123). Our account of simulation is neutral with regard to this distinction between the personal and the subpersonal, since both imagination and attribution have been defined in such way that is silent with respect to the processes and mechanisms

[7] In other words, it is not required for one's imagining having someone else's experiences that one replicates in imagination, either fully or partially, the other person's experiences (Currie 1995; see also Currie 1993; Levinson 1993; Lopes 1998; Morton 2006 on this question). This was also, arguably, Adam Smith's view on sympathy (see Nanay 2010). This being said, when the goal of simulation is epistemological, then replicating faithfully the other person's mental states become the key to a successful simulation.

[8] We have cashed out our characterization of attribution in terms of meta-representation, for it is probably the most familiar way of interpreting attribution. It is not our view however that attribution cannot occur in creatures not capable of, or not yet capable of meta-representation. An agent may be capable of having simultaneous models of how things look for two different people without having representations of representations, and these cognitive abilities might be enough for some forms of attribution (Perner 1993, especially Chaps. 3 and 4). Nothing in this chapter depends on taking a stance on that issue.

in which these capacities are realized, and are therefore compatible with either personal or subpersonal instantiations of them.

The second dimension Morton talks about concerns the degree to which the particular mental processes of the simulated person are taken into consideration by the simulator. At one end of the spectrum, one would engage in what can be called *co-cognition* (Heal 1998, 2000), and at the other end of the spectrum we find *modelling* (Gordon 1995a, b). While the co-cognizer only takes into account the initial challenge the other person is facing, and seeks the answer to this challenge only by means of her own mental resources, the modeller tries to think through the other person's problem in a way that corresponds to the manner in which the modeller believes the other person is likely to think through her problem (Morton 2003, p. 128).

The issue here concerns one of the two notions of perspective we mentioned in connection with imagining from the inside, namely, psychological perspective. When we are simulating, we can attempt to adopt the mental setup of the person we simulate. This might be a good idea if we know that the person we are trying to simulate thinks very differently from the way we do. That is what the modeller does. The co-cognizer, on the other hand, just imagines going through the simulated person's task herself, mobilising the same mental competences she herself would use if she were performing this task. Our account of simulation is compatible with both modelling and co-cognising, as our account is neutral with respect to whether the imagining process that is part of simulation should imply taking into consideration the other person's mental dispositions.

Morton's third dimension, the centred versus non-centred dimension of simulation is again best understood with the help of the notion of perspective: in this case, physical perspective. As we have seen, if I imagine having someone else's experiences, I can do so in two different ways. I can envisage this person's situation entirely from her own physical perspective. However, I can also envisage the other person's situation from the point of view that I occupy. Morton calls the former case 'centred simulation', whereas the latter he calls 'non-centred simulation'.[9] Our characterization covers both centred and non-centred simulation, as our notion of imagination covers both those cases where we take the other person's physical perspective into consideration and those cases where we do not.

Thus, our account of simulation does not rule out any obvious candidate for a theory of simulation expounded in the literature, while it is capable of accommodating the most important ones. The question is whether and in what way this account can illuminate the simulation versus theory-theory debate.

[9] This distinction is similar to the one Wollheim makes between central and acentral imagining (Wollheim 1984; see also Goldie 1999, 2000). It is important to note, however, that Morton's distinction is a distinction between two kinds of simulation, whereas Wollheim's differentiates between two kinds of imagining processes.

20.3 Simulating Versus Theorizing

Simulation involves taking imagining oneself in another person's situation to be a reason for making attributions concerning this person's mental states. We argue that the shape of the simulation versus theory-theory debate depends on the way in which we spell out what 'reason' is taken to mean in this context.

As a first approximation, the difference between simulation and theory-theory could be articulated in the following way: Simulation is attribution that happens as a result of imagination, whereas theory-theory is attribution that happens as a result of applying a theory. In other words: *A* simulates *B* if *A* attributes a mental state to *B* and her reason for doing so is imagining being in *B*'s situation. If *A* attributes a mental state to *B* and her reason for doing so is that it follows from the application of a psychological theory, then *A* is theorizing.

The picture, however, is more complicated. If it is *A*'s application of a psychological theory, rather than her imagining herself in *B*'s situation, that constitutes the reason or justification for attributing a mental state to *B*, then *A*'s attribution falls under the theory-theory model of our mind-reading abilities. If, by contrast, *A*'s imagining herself in *B*'s situation is indeed the reason for *A*'s attribution of a mental state to *B*, then a further question needs to be asked: In what way can a person's imagining constitute a reason for her attribution?

There are many ways in which the idea of reason or justification for attributing mental states can be understood. What is crucial in the present context is whether such justification presupposes the application of a psychological theory. Imagining being in someone else's situation might constitute a reason for an attribution in virtue of the application of a psychological theory, or, alternatively, it might constitute a reason for an attribution without any such theory being applied. However, note that while both of these models qualify as simulation given that in both cases imagining ourselves in the situation of others justifies the attribution of a mental state, the former also looks very much like a version of the theory-theory view in as much as it claims that the attribution of a mental state is (at least partly) justified by the application of a psychological theory.

In order to elaborate on the significance of the distinction between these two versions of simulation, we need to clarify the difference between justification by means of theorizing and justification without any such appeal. These models we call respectively the hybrid model of simulation and pure simulation.

20.4 The Hybrid Simulation Account

The application of a theory might be thought to be best understood with the help of the notion of *inference*. A straightforward way in which imagining can justify attribution in the process of applying the relevant theory is if we *infer* on the basis of imagining ourselves in someone else's situation that we can attribute a certain

mental state to this person. The problem with an appeal to inferences (besides the notorious difficulties surrounding this concept) is the following. Inference may be sufficient for theorizing,[10] but it is much less clear that it is necessary. One could make the case for the claim that, for example, we may engage in theorizing without really drawing any inferences if, for example, the theory in question becomes so entrenched that its application happens quasi-automatically.

Fortunately, we do not need to appeal to the concept of inference to cash out the notion of theorizing at stake here. A weaker requirement on theorizing can be used when trying to differentiate between justification by means of theorizing and justification without theorizing. One such requirement might consist in context-independent representations of the connection between types of imaginings and types of attributions. If it is theorizing in virtue of which imagining justifies attribution, then the agent must have some kind of context-independent representation of the connection between types of imaginings and types of attributions. These representations may be explicit and conscious, as is the case when deliberately drawing inferences. But this is not necessarily the case: they may as well be unconscious representations that nevertheless enable the agent to apply a theory.

Whether inferential and conscious or automatic and nonconscious, if the step from one's imaginings to one's attributions involves the application of theory, we get what we call a hybrid account of simulation. According to the hybrid model of simulation, when *A* imagines being in *B*'s situation and finds herself in a mental state, it is a psychological theory that justifies her attribution of this mental state to *B*. On this picture, simulation and theory-theory would not be two mutually exclusive ways of figuring out someone else's mental state: simulation would presuppose, as a necessary ingredient, the mastery and the application of a theory. Our imaginings can justify our attribution only because of a psychological theory.[11] This would give rise to a hybrid theory-simulation account.

In the last 15 years or so, more and more contributors to the simulation-theory debate have suggested that pure simulation and pure theory-theory should not be thought of as mutually exclusive candidates to account for our mind-reading abilities (Heal 1995, 1998; Goldman 2000; Stone and Davies 1996). The thought is that mind reading does not have to consist in either pure simulation or pure theorizing: it integrates both simulation and theorizing. Note that the hybrid view on offer here echoes this recently popular suggestion of a compromise in as much as we also talk about a way of attributing mental states to others that combines simulation and theorizing. However, what we call the hybrid model consists not merely of a mix of (pure) simulation and (pure) theorizing. The suggestion here is that there is a way of simulating that necessarily involves theorizing since imagination justifies

[10] Even the sufficiency claim is questionable. One possible problem is that while the notion of theory is often taken to presuppose law-like generalizations, inferences can take place in the absence of these.

[11] Note that this interdependence of simulation and theory is very similar to what Davies (1994) suggested (for different reasons) when he examined whether simulation presupposes having tacit knowledge of a psychological theory.

attribution by means of theorizing. In other words, the bonds between simulation and theorizing in the hybrid account are much tighter than it has been suggested.

20.5 The Pure Simulation Account

According to the pure simulationist account, *A* attributes a mental state to *B* as a result of her imagining herself in *B*'s situation without relying on any theory. In other words, *A*'s imagination can justify her attribution of a mental state to *B* without appeal to any theory. Thus, theory would not play any role in simulation; we get a pure (that is, not hybrid) simulationist account. The aim of this section is not to give arguments for the correctness of the pure simulationist account but to show that this is a coherent view and, importantly, to point out how it differs from the hybrid account of simulation.

An analogy with some views in the philosophy of perception may be helpful. Some philosophers argue that perceptual beliefs are justified, not because one uses any theoretical apparatus (of justification), but because perception is a reliable process: *p* reliably causes perceptual beliefs that *p*. Thus, when *p* causes the perceptual belief that *p* reliably (however that is to be understood), *p* in itself is enough to justify having a perceptual belief that *p,* without the use of any theoretical apparatus or even inferences.[12]

The same kind of relation might hold between imagination and attribution. It could be argued that the fact that *A* imagines herself being in mental state *M* in *B*'s situation reliably indicates that *B* is in mental state *M*. But why would imagination be a reliable process? While it makes sense to say that perception is a reliable process (after all, it is a causal one), why claim the same with regard to imagination? The answer is that imagination is not a reliable process in general: If I imagine how my grandchild will look, it is unlikely that I will get it right. But a special case of imagination, namely, imagining being in someone else's situation, may reliably indicate the other person's mental states, provided that the mental setup of the simulator and that of the simulated person are sufficiently similar.[13] Thus, we can say that *A*'s imagining herself being in mental state *M* in *B*'s situation justifies *A*'s attribution of *M* to *B* without any appeal to theory. If this is the case, then it is possible to take one's imagination to be the reason for one's attribution without the use of any theory.

[12] Reliabilism is one way in which the idea we are after can be cashed out (see, e.g. Goldman 1967, 1976; Plantinga 1993; Dretske 1981; Nozick 1981; Swain 1981; Armstrong 1973). We are not interested in adjudicating between different versions of reliabilism or to take sides with any view in epistemology or in the philosophy of perception. The role of this analogy is to point out that there are perfectly consistent and even fashionable views regarding justification that do not appeal to the application of any theory.

[13] See Morton's observations (Morton 2006) about the circumstances under which imagining someone else is likely to be successful.

An example may be useful. Gordon (1995a, 1996) has attempted to articulate such a non-hybrid view under the label 'radical simulation'. According to Gordon, at least a subset of simulation episodes are cases of simply recentering oneself in the simulated person's location and concerns through what he calls 'an egocentric shift' (Gordon 1995a, p. 56). On this picture, one simply sees the situation from the perspective of the other and registers how the world is for this other person. The distinguishing feature of this way of adopting someone else's perspective on the world, Gordon argues, is that it does not require the mastery and the application of any mental concepts on behalf of the simulator and thus excludes the possibility that simulating presupposes an episode of theorizing. Although the idea of undergoing an 'egocentric shift' is a notoriously difficult one to cash out (see Heal 1995, p. 44), from our perspective, what is important is that Gordon's theory, whatever its merits, falls under the category that we labelled 'pure simulation' (see also Jeannerod and Pacherie 2004).[14]

We have considered two ways in which imagination can justify attribution. If this justification happens with the help of a theory, then we get hybrid simulation. If, on the other hand, the justification does not require any theory, then we face pure simulation.

20.6 Conclusion: An Epistemological Turn?

The aim of this chapter was to show that the simulation versus theory-theory debate is in a large part epistemological. We found that we need to distinguish three different ways in which one could be said to justifiably attribute a mental state to someone else:

1. Imagining oneself in the other person's situation justifies the attribution of mental states to her without recourse and deployment of a psychological theory (pure simulation).
2. Imagining oneself in the other person's situation justifies the attribution of mental states to her with recourse and deployment of a psychological theory (hybrid simulation).
3. The attribution of mental states is justified without any use of imagination (theory-theory).

The debate regarding which of these three is the right way to think about mental attribution in the present context is an epistemological one in at least three distinct senses.

[14] Some are persuaded that pure simulation finds at least some confirmation by recent neuroanatomical findings (Gallese and Goldman 1998; Adams 2001, for example). The suggestion is that mirror neurons provide the physiological basis for simulation. If this suggestion is correct, then there is a way of simulating other people that does not require the use of any theory. Thus, this proposal is also a version of the 'pure simulation' account. It is not our aim here to evaluate this proposal, but to point out that this would count as an example of a 'pure simulation' view.

First, general epistemological considerations put constraints on the simulation versus theory-theory debate. Depending on what epistemological background theory one has, the simulation versus theory-theory debate will look very different. If one denies the possibility of justification without the deployment of at least some theoretical apparatus, then (1) is not an option: the debate will be between (2) and (3), both of which involves some reference to theories. If, on the other hand, one favours something like reliabilist accounts of justification and thus allows for justification without the deployment of any theory, then the debate will revolve around the question of whether it is possible that imagination justifies attribution without any use of a theory. In that case the real divide will be between (1) on the one hand and (2) and (3) on the other. In short, one's general views in epistemology will determine the shape of the simulation versus theory-theory debate.

Second, we aimed to point out that all three accounts of mind-reading appeal to some notion of justification. Hence, they could not be fully spelled out without using one or another specific notion of justification. Depending on what notion of justification we plug in into these accounts, we will get different theories of mind-reading with varying plausibility.

Third, the stance we take with regard to the simulation versus theory-theory debate put constraints on one's general epistemological considerations. Depending on what account of mind-reading one finds plausible, one ends up with different implications with regard to more general epistemological claims. If for example one finds it plausible that our imagining episode can serve as a reason for attributing mental states to others without relying on any theoretical apparatus this commits one to a version of something like reliabilism about justification. Thus, when adjudicating between the accounts of mind-reading we differentiated, the epistemological assumptions behind these accounts should not be ignored.

Let us conclude by observing that the claim we are making is not that the simulation versus theory-theory debate is *just* an epistemological debate and nothing more. Even if we resolved all general epistemological disagreements, this would not settle the simulation versus theory-theory debate. This debate is partly psychological and we do not want to deny this. Our contention is that it is *partly* psychological, but it is also partly epistemological: Psychology alone, without bringing in epistemological considerations, will not settle this debate. In fact, the failure to distinguish clearly the epistemological issues discussed here from the psychological ones might have contributed to the confusion of the various dimensions of this debate.

Thus, we suggest an epistemological turn in the simulation versus theory-theory debate: in order to settle, or even to take a stance on, this issue, we need to bring in epistemological considerations. But this does not mean that we should forget about the psychological nature of this debate. The suggested epistemological turn in the simulation versus theory-theory debate is really only a half turn.

References

Adams F (2001) Empathy, neural imaging and the theory versus simulation debate. Mind Lang 16:468–392
Armstrong DM (1973) Belief, truth and knowledge. Cambridge University Press, Cambridge
Currie G (1993) Impersonal imagining, a reply to Jerrold Levinson. Philos Quart 43:79–82
Currie G (1995) Image and mind: film, philosophy, and cognitive science. Cambridge University Press, Cambridge
Davies M (1994) The mental simulation debate. In: Peacocke C (ed) Objectivity, simulation and the unity of consciousness. Current issues in the philosophy of mind, Proceedings of the British Academy 83. Oxford, Oxford University Press, pp 99–127
Dretske F (1981) Knowledge and the flow of information. The MIT Press, Cambridge
Gallagher S (2005) How the body shapes the mind. Oxford University Press, Oxford
Gallagher S (2007a) Logical and phenomenological arguments against simulation theory. In: Hutto D, Ratcliffe M (eds) Folk psychology re-assessed. Springer, Dordrecht, pp 63–78
Gallagher S (2007b) Simulation trouble. Soc Neurosci 2:353–365
Gallese V, Goldman A (1998) Mirror neurons and the simulation theory of mind-reading. Trends Cogn Sci 3:493–501
Goldie P (1999) Understanding other people's emotions. Mind Lang 14:394–423
Goldie P (2000) The emotions: a philosophical exploration. Oxford University Press, Oxford
Goldie P (2001) Emotion, personality and simulation. In: Goldie P (ed) Understanding emotions: mind and morals. Ashgate Publishing, Aldershot
Goldman A (1967) A causal theory of knowing. J Philos 64:357–372
Goldman A (1976) Discrimination and perceptual knowledge. J Philos 73:771–791
Goldman A (2000) The mentalizing folk. In: Sperber D (ed) Metarepresentations. Oxford University Press, Oxford
Gordon RM (1995a) Simulation without introspection or inference from me to you. In: Davies M, Stone T (eds) Mental simulation. Blackwell, Oxford, pp 53–67
Gordon RM (1995b) Sympathy, simulation, and the impartial spectator. Ethics 105:727–742
Gordon RM (1996) 'Radical' simulationism. In: Carruthers P, Smith PK (eds) Theories of theories of mind. Cambridge University Press, Cambridge, pp 11–21
Heal J (1995) How to think about thinking. In: Davies M, Stone T (eds) Mental simulation. Blackwell, Oxford, pp 33–52
Heal JB (1998) Co-cognition and off-line simulation: two ways of understanding the simulation approach. Mind Lang 13:477–498
Heal JB (2000) Other minds, rationality and analogy. Aristot Soc Suppl 74:1–19
Hutto DD (2007) Folk psychology without theory or simulation. In: Hutto D, Ratcliffe M (eds) Folk psychology re-assessed. Springer, Dordrecht, pp 115–135
Hutto DD (2008) Folk psychological narratives: the sociocultural basis of understanding reasons. MIT Press, Cambridge
Jeannerod M, Pacherie E (2004) Agency, simulation and self-identification. Mind Lang 19, 113–146
Levinson J (1993) Seeing, imaginarily, at the movies. Philos Quart 43:70–78
Lopes DM (1998) Imagination, illusion and experience in film. Philos Stud 89:343–353
Morton A (2003) The importance of being understood: folk psychology as ethics. Routledge, London
Morton A (2006) Imagination and misimagination. In: Nichols S (ed) The architecture of the imagination. Oxford University Press, Oxford
Nanay B (2010) Adam Smith's concept of sympathy and its contemporary interpretations. Adam Smith review 5:85-105. Reprinted. In: Brown V, Fleischacker S (eds) The philosophy of Adam Smith. Routledge, London, pp 85–105
Nozick R (1981) Philosophical explanations. Harvard University Press, Cambridge
Perner J (1993) Understanding the representational mind. The MIT Press, Cambridge

Plantinga A (1993) Warrant: the current debate. Oxford University Press, Oxford
Ratcliffe MJ (2007) Rethinking commonsense psychology: a critique of folk psychology. Theory of mind and simulation. Palgrave Macmillan, Basingstoke
Smith M (1995) Engaging characters. Oxford University Press, Oxford
Smith M (1997) Imagining from the inside. In: Allen R, Smith M (eds) Film theory and philosophy. Oxford University Press, Oxford, pp 412–430
Stone T, Davies M (1996) The mental simulation debate: a progress report. In: Carruthers P, Smith PK (eds) Theories of theories of mind. Cambridge University Press, Cambridge, pp 119–137
Swain M (1981) Reasons and knowledge. Cornell University Press, Ithaca
Walton K (1990) Mimesis and make-believe. On the foundations of the representational arts. Harvard University Press, Cambridge
Wollheim R (1974) Identification and imagination. In: Wollheim R (ed) Freud: a collection of critical essays. Anchor Press, New York, pp 172–195
Wollheim R (1984) The thread of life. Harvard University Press, Cambridge
Wollheim R (1987) Painting as art. Thames and Hudson, London
Zahavi D (2008) Simulation, projection and empathy. Conscious Cogn 17:514–522

Chapter 21
Mental Simulation and the Reification of Beliefs

Jérôme Dokic

Abstract Simulation theory has been put forward as an account of our folk understanding of the mind. In this chapter, I examine a neglected argument to the effect that there is an essential limitation of simulation itself, which cannot explain a crucial ingredient of our ordinary, folk-psychological conception of beliefs. Even if it is conceded that simulation gives the subject some sense of what happens in the world when someone believes something, the understanding of *facts* of believing that can be extracted from simulation is incomplete; simulation must be augmented with a theory of beliefs as genuine *constituents* of such facts. Folk psychology reifies beliefs in order to deal with an essential requirement for mastery of the folk-psychological concept of belief. Hopefully, a reflection on the limits of simulation will lead to a better understanding of the role of theory in ordinary belief-ascriptions.

Keywords Mental simulation · Theory of mind · Ontology of belief · Reification of beliefs · False belief task

21.1 Introduction

In the last few decades, philosophers of mind and cognitive psychologists have debated about what our ordinary understanding of the mind consists in. The original discussion concerned the issue of whether this understanding should be explained in terms of our mastering a theory, or in terms of a general capacity for simulating others. According to the so-called *theory-theory*, we posit unobservable entities (mainly beliefs and desires) in order to explain observed facts (mainly behaviour), using our mastery of psychological laws. According to the rival, *simulation theory,* we understand others by trying to get our own mind to work in relevant ways like

J. Dokic (✉)
Institut Jean-Nicod (CNRS, EHESS, ENS), Paris, France
e-mail: dokic@ehess.fr

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_21,

theirs. Simulation in this sense is a practical ability, which does not require a sophisticated grasp of psychological laws.[1]

What has emerged from this debate is that folk psychology probably uses both strategies to a certain extent, depending on what is relevant in the context. However, it is fair to say that the respective contributions of simulation and theory are still not entirely clear. In this chapter, I would like to examine a neglected argument to the effect that there is an essential limitation on simulation, which cannot explain a crucial ingredient of our ordinary conception of belief. Even if it is conceded that simulation gives the subject some sense of what happens in the world when someone believes something, the understanding of *facts* of believing that can be extracted from simulation is incomplete—simulation must be augmented with a theory of beliefs as genuine *constituents* of such facts. Hopefully, the following discussion will lead to a better understanding of what is involved in ordinary belief-ascriptions (namely statements of the form "*S* believes that *p*", where *S* is a subject and *p* a truth-evaluable proposition).

I should make clear at the outset that I am concerned here with the issue of what our ability to *ascribe* mental states consists in. In other words, I am interested in our *theoretical* understanding of the mind, and the extent to which such understanding also depends on practical simulation. Now, one can argue that our understanding of the mind can also be purely practical, i.e. characterized by "action, involvement and interaction based on environmental and contextual factors" rather than as "explanation or prediction based on mental contents" (Gallagher 2005, p. 212). In the present chapter, I have nothing to say about the possibility of a purely practical understanding of the mind, which arguably does not involve the explicit ascription of mental states as such.

This chapter is structured as follows. In the next three sections, I formulate what seems to me to be the most convincing version of simulation theory as an account of belief-ascriptions, which I call "the quasi-modal account of beliefs". Although it owes much to Robert Gordon's work, it is less radical than his own simulation theory, since an important objection to Gordon's theory has been taken into account. In Sect. 21.5, I try to show that simulation theory is unable to deal with an essential requirement for mastery of the concept of belief. In Sect. 21.6, I strengthen my argument by considering an analogy with tense logic. In Sect. 21.7, I suggest that folk psychology meets the relevant requirement by referring to, or quantifying over beliefs, thus going beyond the quasi-modal account. In the concluding section, I ask whether the argument of this chapter shows that the quasi-modal account is wrong, or just incomplete. I am inclined to favour the latter option.

[1] Many of the relevant papers are collected in Davies and Stone (1995a, b) and Carruthers and Smith (1996). See also Dokic and Proust (2002) and Goldman (2006).

21.2 Mental Simulation

The main tenet of simulation theory, as an account of our folk understanding of mental states, is that the ability to *imagine* or *simulate* worlds possibly different from the actual one can be exploited to adopt alien perspectives—to represent the world not only as I have found it, but as others have found it, or as I might have found it.

Sometimes, the claim is that merely imagining the world from the other's perspective is enough to ascribe the relevant beliefs to her:

> to attribute a belief to another person is to make an assertion, to state something as a fact, *within the context of practical simulation*. Acquisition of the capacity to attribute beliefs is acquisition of the capacity to make assertions in such a context. (Gordon 1995a, p. 68)

Indeed, Gordon often comes close to what might be called "the non-conceptual view of belief-ascription", according to which the capacity to ascribe beliefs does not require antecedent possession of psychological concepts. If practical simulation can ground belief-ascription, and does not require mastery of the concept of belief, belief-ascription itself does not have to deploy such a concept.

However, as many critics of Gordon have pointed out, simulating the world, even as the other has found it, falls short of ascribing a belief (Heal 1994, p. 136; Jacob 2002; Currie and Ravenscroft 2002). It is one thing to *engage* in simulation, but it is another thing to *describe* or *exploit,* in a detached way, the results of simulation. If simulation is to underlie ascriptions of mental states, the subject should be able to understand the specific consequences that it has beyond its scope, and in particular on what is the case in the real world as she conceives it. What must be explained is the significance of *embedding* a particular simulation within her own *doxastic* perspective, i.e. the perspective of her beliefs about the world.

This much should be conceded to theory-theory: The mere capacity to engage in practical simulation does not give the subject an idea of what is the case in the real world when someone believes something. This is true even if the simulation happens to be successful, i.e. if the simulator actually gets her mind to work like the simulatee's.

However, the nonconceptual view of belief-ascription is not the only option available to someone who wants to give simulation a central role in a theory of how we understand others and ourselves. Even if the exploitation of simulation in belief-ascription essentially involves psychological concepts, the latter may not include the concept of belief as a theoretical posit, as theory-theory claims. In other words, two questions should be distinguished:

1. Can simulation ground belief-ascription without deploying any psychological concepts?
2. Can simulation ground belief-ascription without deploying the concept of belief as a mental state?

What I have claimed so far is that we should give a negative answer to the first question. A nonconceptual view of belief-ascription is hopeless. However, this leaves

the second question open. Theory-theory also gives a negative answer to it. Now, perhaps a simulative account of belief-ascription can show that theory-theory over-intellectualises beliefs in this respect.

21.3 The Quasi-Modal Account of Beliefs

Compare mental simulation with games of make-believe. A child who pretends that a particular banana is a telephone needs at least the capacity to keep in mind several "situations" or "mental models" at once (Perner 1991). In one model, she is holding a banana while in the other model she is holding a telephone. The representations "I am holding a banana" and "I am holding a telephone" do not conflict with each other because they belong to different mental models. The former representation describes her doxastic world, i.e. the world as she believes it to be, whereas the latter representation describes her imaginary world, i.e. the world as she imagines it to be.

Strictly speaking, the capacity to engage in games of make-believe does not require an explicit representation of the pretence as such. There is an analogy with desire here (Currie 1998; Nichols and Stich 2000). Just as we can imagine a creature that acts on its desires without having any belief of the form "I desire that *p*", we can imagine a creature that plays games of make-believe without having any belief of the form "I pretend that *p*".

However, a minimally reflective creature will have at least beliefs, or experiences, about what is desirable (from its point of view). Similarly, a minimally reflective creature will have some beliefs about the pretence. For instance, if our child is asked what she is doing, she might be old enough to give the following answer:

(1) I am holding a banana, but in the game, I am holding a telephone.

The representation "In the game, I am holding a telephone" is a natural expression of the *embedding* of her imaginary model within her doxastic one. This complex representation, which is part of the child's model of reality, contains another representation ("I am holding a telephone"), originally produced within the child's imaginary model and now indexed to a particular game of make-believe.

Consider now mental simulation. Just as one can index an imaginary situation to a particular game of make-believe, perhaps one can index an imaginary situation to a particular person. Even though I believe that it is raining, I can imagine the world as it would be if it were sunny. Furthermore, I can imagine such a world from the perspective of a particular person, for instance Pierre. I can thereby form a complex representation such as the following:

(2) It's raining, but according to Pierre, it's sunny.

The representation "According to Pierre, it's sunny" is a natural expression of the embedding of my imaginary model within my doxastic one. The expression "according to Pierre" is, like "in the game", an intensional operator modifying a representation which the subject may not consider true. Of course, much more has to

be said about such operators if we are to explain the significance of embedding one mental model within another. However, it is worth observing that intuitively, "According to Pierre, it's sunny" is true if and only if my simulation of Pierre is reliable (I am getting my mind to work like Pierre's), and the latter is reliable if and only if Pierre *believes* that it is sunny.

On what I shall call "the quasi-modal account of beliefs", the phrase "Pierre believes that" has more or less the same sense as "according to Pierre". Prior and others have defended the view that this phrase is a unary sentential operator, on a par with the operators used in modal logic (Prior 1963; and, for a sympathetic account of Prior's view, Recanati 2000, § 3.2). What is added here is the idea that its use is based on mental simulation. I understand "Pierre believes that *p*" as "According to Pierre, *p*", namely by running and exploiting a simulation of Pierre. Just as modal operators like "it is possible that" and "it is necessary that" do not explicitly refer to or quantify over possible worlds, judgements of the form "According to Pierre, it's sunny" do not represent doxastic perspectives or beliefs as such. On the quasi-modal account, belief-ascriptions deploy psychological concepts, namely those expressed by "according to *X*", but do not introduce beliefs as *objects* (in the broad Fregean sense of the term, which includes abstract entities). The quasi-modal account of beliefs is ontologically *neutral* in this respect.

21.4 Nested Simulations

On Gordon's view, the ability to simulate the world from the other's perspective gives sense to the idea of a "mental location", i.e. "to the notion of something's being a fact *to* a particular individual" (1996, p. 18). It could be argued that a representation such as "According to Pierre, it's sunny" gives expression to this idea; I can use it to represent that it is a fact *to* Pierre that it is sunny although I myself believe that it is raining.

However, Gordon makes clear that a subject who imagines the world from the other's perspective, even if it is incompatible with her own perspective, still falls short of manifesting possession of the ordinary concept of belief. The problem is that she "is not yet in a position to understand that *her own present beliefs* may themselves deviate from the facts". Indeed, "she will not come to understand this as long as she ascertains what her own present beliefs are by asking what the facts are" (1995b, p. 62). Now what is remarkable is that Gordon thinks that an appropriate *nesting* or *embedding* of simulations can yield the more sophisticated notion of belief:

> To see her own present beliefs as distinguishable from the facts she will have to simulate another for whom the facts are different—or, more broadly, adopt a perspective from which the facts are different, whether this perspective is occupied by a real person or not—and then, from the alien perspective, *simulate herself.* (1995b, p. 62)

Gordon describes another way in which, according to him, one can come to understand that one's present worldview can be wrong. It can arise with "the capacity to demote some of one's own *memories* to off-line status" (1995b, p. 62). I seem to remember that this was a rock but I now know that it is a sponge. I come to understand that my previous worldview was wrong:

> by simulating a possible later perspective of her own the child may come to conceptualize in a more sophisticated way what she now counts naïvely as fact pure and simple. She comes to think of it instead as fact relative to her present perspective. (1995b, p. 62)

Now, what Gordon says about embedded simulations also applies to the representations that give expression to such embeddings. In effect, the claim is that one has to understand that the operator "according to *X*" is *recursive,* so that one can form sophisticated representations such as the following:

> (3) *p*, and according to someone (¬*p* and according to me (*p*)).

The idea is that I must go through a representation of someone else, or of myself at a different time, to access my present beliefs conceived as one fallible perspective among others. Thus, on the quasi-modal account of beliefs, my understanding of false beliefs can be fully manifest in my use of "according to *X*" as a recursive operator.

21.5 The Limits of Simulation

Does the quasi-modal account of beliefs really succeed in dealing with belief-ascriptions? What Moore's paradox shows (among other things) is that although there is something odd about a statement like "I believe that it's raining, but it's not raining", it is not a logical contradiction. I might believe something false. Indeed, it is a minimal requirement for mastery of the concept of belief that one can make sense of the metaphysical possibility that I believe that *p* while ¬*p:*

> *Minimal Requirement (for mastery of the concept of belief):* Anyone who masters the concept of belief can conceive of the truth of "Possibly, I believe that *p* but ¬*p*" for at least some contingent propositions *p*.[2]

Independently of whether the competent subject reifies beliefs, she should be able to understand that the fact that she believes that *p* is consistent with ¬*p* being a fact, or equivalently with *p* being a mere (i.e. nonfactual) state of affairs.

What I would like to show is that the quasi-modal account of beliefs goes no way toward explaining how the Minimal Requirement can be met. Let us consider again the kind of representation that according to the quasi-modal account manifests our understanding of the notion of false belief:

> (4) It's raining, but according to Pierre (it's sunny, and according to me, it's raining).

[2] It has been objected (by Josh Mozersky) that this formulation does not allow for an essentially omniscient being. My answer is that an essentially omniscient would not have *beliefs,* and so would not use folk psychology as we know it.

The fact that a subject makes sense of (4) does not yet show that she understands that the embedded representation, "It's sunny, and according to me, it's raining", expresses a metaphysical *possibility*. Note that in general, the quasi-modal account must deal with the ascription of beliefs with impossible contents, such as the following:

(5) Hesperus is Phosphorus, but according to Pierre, Hesperus is not Phosphorus.
The square root of 4 is 2, but according to Pierre, the square root of 4 is 3.

Perhaps one can understand the representation "According to Pierre, Hesperus is not Phosphorus" by running a simulation in which there seem to be two planets, but it does not follow that the simulator thinks that there is a possible world in which Hesperus is not Phosphorus. On the contrary, she pictures Pierre's world as an *impossible* one. Similarly, I can simulate a world in which it seems that the square root of 4 is 3 (for instance, I simulate Pierre doing a calculation whose result is that the square root of 4 is 3), but it by no way follows that I picture Pierre's world as a possible one. Now the point is that (4) could be in the same boat as the examples in (5). The mere fact that the representation "It's sunny, and according to me, it's raining" is used in the scope of the quasi-modal operator "according to Pierre" does not entail that Pierre's world is a possible one *according to the subject*.

In other words, it is far from obvious that embedding quasi-modal operators can yield an understanding of the notion of false belief. Such embedding can at best yield an understanding that others *think* that one is actually wrong; it cannot yield an understanding that others are *possibly right* in thinking that one is wrong. Similarly, the capacity to demote some of one's own memories to offline status, as well as the capacity to simulate a possible later perspective of one's own, are insufficient to ground mastery of the notion of belief. On the quasi-modal account of beliefs, a subject can conceive of a later perspective according to which she is now wrong, but this is not how she understands the possibility that such a perspective is the correct one. By embedding appropriate quasi-modal operators, she can give expression to the idea that she was wrong in the past but not that she might have been right then. In general, the subject cannot express in purely quasi-modal terms the possibility that one of her own beliefs is false, *independently of anyone's opinion on the matter*. The Minimal Requirement is left unexplained.

21.6 An Analogy with Tense Logic

An analogy with tense logic might be useful at this point. Tense logic in the modern form invented by Prior involves "tensed" sentences which are true or false only relative to a time. Tensed sentences can be uttered on their own, or embedded in temporal operators, such as "in the past" and "in the future". When a tensed sentence like "Brown is ill" is uttered on its own, it is understood as "implicitly characterizing" the time of the utterance, so that the sentence is true if and only if Brown is ill at that time. The sentence can also characterize other times when it is used in the

scope of some temporal operator. For instance, an utterance of "In the past, Brown is ill" is understood as characterizing some time before the time of the utterance.

Now as Prior was perfectly aware, this introduces an *asymmetry* between the time which is *present* from the speaker's point of view and other times:

> It is clear that although the […] tensed language *mentions* no instants there is a sense in which it *implicitly refers* to the time of utterance, and by tensing what is implicitly said of the time of utterance it can indirectly characterise other times too, also these are referred to rather indefinitely. […] But every complete tensed sentence characterises the time of utterance in some way or other, and *other times only through their relation to that one.* (Prior, 2003, pp. 224–225; last italics mine)

In its simplest form, tense logic makes "now" redundant in the sense that whenever it is correct to assert a tensed sentence on its own, it is correct to assert this sentence modified by "now". At some point, Prior was tempted to go further, and defend what he calls "a redundancy theory of the present tense", which he formulates as follows:

> I have argued that, whatever the proposition that *p* might be, the proposition that *it is (now) the case that p* is the very same proposition as the proposition that *p*. For instance, the proposition *that it is now the case that I am sitting down* is the very same proposition as the proposition *that I am sitting down* […]. (2003, p. 171)

However, Prior quickly realized that there are counterexamples to the redundancy theory of the present tense. As is now well known, the statement "In the future, Brown is ill" is not equivalent to the statement "In the future, it is now the case that Brown is ill". The former statement is true if and only if Brown will be ill (relative to the time of utterance), whereas the latter statement is true if and only if it will be true that Brown is ill at the time of utterance. In David Kaplan's words, "now" cannot be controlled by an operator, for it will simply "leap out of its scope to the front of the operator" (1989, p. 510). A plausible explanation of this phenomenon is that "now" directly *refers* to the time of utterance. This explanation puts in jeopardy the ontological neutrality of tense logic.

Now, consider the following analogy between tense logic and the quasi-modal account of beliefs. In the latter account, there is also an asymmetry, between one's own doxastic perspective and others'. When a sentence is used on its own, it implicitly characterizes the subject's perspective, at least in the sense that she sincerely asserts *p* if and only if her doxastic world includes the fact that *p*. As a consequence, whenever it is correct to assert *p,* it is correct to assert "According to me, *p*". The quasi-modal operator "according to me" corresponds to the adverb "now" in the temporal case. Other doxastic perspectives are characterized only through their relation to the subject's, for instance as "According to *X*, it's sunny", where *X* is conceived to be someone else.

It does not follow that "*p*" and "According to me, *p*" can be substituted *salva veritate* in any context, at least if the latter is intended to have the same sense as "I believe that *p*". In particular, the statement "According to Pierre (it is sunny and according to me, it is raining)" is not equivalent to "According to Pierre (it is sunny and it is raining)". In contrast to the first statement, the second ascribes to

Pierre a belief with an impossible content.[3] Similar remarks can be made about the behaviour of "according to me" in the scope of modal operators. For instance, the true statement "It is possible that (it is sunny and according to me, it is raining)" is not equivalent to the necessarily false statement "It is possible that (it is sunny and it is raining)".

The friend of tense logic must meet the challenge of explaining the behaviour of "now" in the scope of modal and quasi-modal operators while staying ontologically neutral as far as times are concerned. The point of the analogy is that the friend of the quasi-modal account of beliefs must also meet the challenge of explaining the behaviour of "according to me" in the scope of modal and quasi-modal operators while staying ontologically neutral as far as doxastic perspectives are concerned.

Perhaps, a sophisticated simulationist will find a way of meeting this challenge, and explaining the Minimal Requirement within the quasi-modal account of beliefs. For my part I have been unable to think of any such way. In the rest of this chapter, I shall rather try to show that folk psychology goes beyond the quasi-modal account of beliefs in order to deal with the Minimal Requirement.

21.7 The Reification Argument

According to what I shall call "the Reification Argument", the Minimal Requirement is explained by *reference* to or *quantification* over doxastic perspectives or beliefs. We ordinarily conceive the fact that someone believes something in terms of the existence of a mental state, belief, which is posited as a real object. In other words, folk psychology *reifies* beliefs:

Reification: S believes that $p \rightarrow$ There is a belief b such that b is true iff p.

Reification of beliefs is what allows us to place what we see as a true belief in a possible world in which it is false. In particular, it allows us to make sense of the following statement:

(6) My belief that p is true, but it is metaphysically possible that it is false.

This statement is *de re* relative to my current doxastic perspective, since the anaphor "it" in the scope of the modal operator "it is possible that" points back to the nominalization "my belief that p", which is outside the scope of the operator. Thus, the truth of "It is possible that I believe that p while $\neg p$" (which requires, for instance, the possible coexistence of the fact that I believe that it is raining and the fact that it is sunny) is understood in terms of the independence of a *belief* from what makes it true or false in the world. In (6), I identify my belief *across* possible worlds, thus expressing the idea that the same belief can be true in one world and false in another. Reification explains the Minimal Requirement.

[3] I presuppose that a single place is in question, and that the feature sunny excludes the feature rainy.

The claim that folk psychology reifies mental states such as beliefs has a surprising implication for the interpretation of well-known empirical data in cognitive psychology. Following an initial suggestion by Dennett (1978), so-called false belief tasks have been designed to determine whether children have the ability to attribute false beliefs to others. The classical version of these tasks, due to the developmental psychologists Hans Wimmer and Josef Perner (Wimmer and Perner 1983), involves a character called "Maxi". Maxi hides a chocolate bar in a blue box, and then leaves the room. In Maxi's absence, another character moves the chocolate bar from the blue box to another, red box. Maxi returns, and the child is asked to indicate where Maxi is likely to look for the chocolate bar. The well-known results are that children around the age of three tend to indicate the red box (where the chocolate bar really is), while older children tend to indicate the blue box (where Maxi falsely believes the chocolate bar to be). Wimmer and Perner's conclusion was that only the older children, around the age of four, master the concept of false belief.

In recent years, many psychologists have objected that the classical version of the false belief task is too intellectualistic. Indeed, it seems to be quite difficult for the children even independently of the requirement to reason about false belief. So, 3-year-olds might fail the false belief task because of general task demands (Bloom and German 2000). Other versions have been designed, including nonverbal ones, that seem to show that children much younger than three, including babies, already have some capacity to understand that others can have false beliefs (Onishi and Baillargeon 2005; Surian et al. 2007; Baillargeon et al. 2010).

If the argument of the present chapter is on the right track, most of these experiments, including the original versions by Wimmer and Perner, do not really test the mastery of the folk-psychological concept of belief. Insofar as they involve the ascriptions of mental states to *others,* no conclusion can be drawn about whether children understand that one of their own present beliefs might be false. Again, a quasi-modal conception of beliefs is enough to understand that someone else has a wrong picture of reality, in the sense of having a set of beliefs that are at odds with one's own picture of reality. Genuine false beliefs tasks should have essentially a self-referential component, which is lacking in many tasks described in the current literature.

So, what do these versions test? One possible answer is that some of them test the presence of a form of purely practical understanding of the mind, which does not involve the ascription of mental states as such. Another answer is that they test the presence of a conception of a type of mental states less sophisticated than beliefs, for instance some epistemic or proto-doxastic relation to, rather than a separate doxastic representation of, the external world.[4] In order to understand such an epistemic relation, the child should be able to distinguish a situation in which the relation holds, and a situation in which it does not hold. In contrast, understanding the folk-psychological concept of belief requires being able to distinguish three types

[4] See Bartsch and Wellman (1995), who draw on both conceptual and developmental grounds a distinction between understanding mental states as mere connections and understanding them as genuine representations.

of situation: a situation in which the other has a true belief, a situation in which she has a false belief, and a situation in which she does not hold the belief at all. The former requires only a quasi-modal conception of mental states, whereas the latter involves the reification of beliefs as mental entities ontologically distinct from the state of affairs they are about.

Let me conclude this section with a few points of clarification. First, the Reification Argument is independent of the issue of whether beliefs should be conceived as concrete objects, like brain states, or as more abstract ones. Beliefs are objects that one cannot easily count, and it has been argued (notably by Steward 1997) that the type-token distinction, which has a point in the case of concrete objects, does not apply to beliefs, just as it does not apply to numbers. Perhaps folk psychology itself is neutral on this issue.

Second, the Reification Argument does not entail that there is a source of direct knowledge of beliefs, such as introspection. I take Frege to have shown, in opposition to an influential Kantian tradition, that objects can be given independently of a source of (internal or external) knowledge. Frege thought, perhaps wrongly, that numbers are such objects. One goes beyond the Reification Argument if one claims that beliefs are like numbers in this respect.

Third, the Reification Argument leaves open the possibility of further analysis of beliefs, for instance as involving relations to propositions. I am not sure that this analysis is the correct one, but it is at least prima facie consistent with the picture of beliefs as entities whose existence or instantiation can be independent of what makes them true or false in the world.

Finally, the Reification Argument does not rest on purely linguistic grounds. It is plausible that the verb "believe" is grammatically prior to the nominalization "belief". It might then be argued that such nominalization is harmless and does not come with genuine reification of beliefs.[5] In contrast, the Reification Argument shows that reification is needed to make sense of our ordinary concept of belief. It is a genuine ontological step in our conception of facts of believing.

21.8 Conclusion

According to the main argument of this chapter, a language that can be used to ascribe full-blown beliefs should have enough expressive power to refer to, or quantify over them. Arguably, this means that we have to make an important concession to "theory-theory": Our ordinary concept of belief as a propositional attitude cannot be extracted from simulation alone but results from a theory of mind that posits beliefs as (abstract or concrete) *objects*. Simulation is not in itself metarepresentational, and modal constructions of the form "According to Pierre, it's sunny" are metarepresentational only in a weak sense: They are representations involving other representations as semantic proper parts (Bermúdez 2003, § 9.4). In contrast, ascriptions

[5] Friederike Moltmann raised this objection to me about an earlier version of this chapter.

of beliefs such as "Pierre believes that it's sunny" involve, at least tacitly, the identification of a mental representation as such. Thus, they are metarepresentational in a strong sense, insofar as they explicitly involve an ontology of mental objects.

Does it follow that simulation fails to play any role in an account of our capacity to ascribe beliefs to others? Everyone agrees, of course, that simulation can play an important *epistemological* role, if it is set against the appropriate theoretical background. If I succeed in getting my mind to work like Pierre's, i.e. if my simulation of Pierre is reliable, I have a warrant for my ascription to him of the belief that it is sunny. The question is rather whether simulation plays any role in our *understanding* of belief-ascriptions.

As far as the argument of this chapter is concerned, there seem to be two alternatives. On the first alternative, the notion of simulation should simply be banned from the account of our understanding of belief-ascriptions. Following theory-theory, such understanding is pictured as relying on our mastery of psychological laws. These laws relate objects, namely beliefs, whose existence is independently acknowledged by the Reification Argument.

The second alternative is, I think, more interesting and promising. The Reification Argument does not show that the quasi-modal account of beliefs is wrong. Rather, it shows that it is incomplete, and must be augmented with an appropriate reflection on the nature of beliefs. On the second alternative, the quasi-modal account of beliefs yields some understanding of *facts* of believing, and the Reification Argument shows that such facts must be further analysed as involving mental states as *constituents*. A virtue of this hybrid or two-tiered account of our understanding of belief-ascriptions is that although beliefs are posited as objects, an important insight of simulation theory is taken on board: Our understanding of the beliefs of others and the rational connections between them is not based on explicit consideration of psychological laws. Rather, as Jane Heal (1996, 1998) has suggested, the connections between beliefs follow the connections between facts in the imagined world of the other.

Acknowledgments Ancestors of this chapter have been presented at the Institute Jean-Nicod in Paris and at Queen's University in Kingston. I thank Dick Carter, Eros Corazza, Elisabeth Pacherie, Adèle Mercier, Friederike Moltmann and Josh Mozersky for helpful comments. This chapter like almost anything else I have written owes much to Kevin Mulligan, of course, who opened my eyes to what genuine philosophy is (which unfortunately does not mean that I am able to reproduce it myself).

References

Baillargeon R, Scott RM, He Z (2010) False-belief understanding in infants. Trends Cogn Sci 14(3):110–118

Bartsch K, Wellman HM (1995) Children talk about the mind. Oxford University Press, Oxford

Bermúdez JL (2003) Thinking without words. Oxford University Press, Oxford

Bloom P, German TP (2000) Two reasons to abandon the false belief task as a test of theory of mind. Cognition 77:B25–B31

Carruthers P, Smith PK (eds) (1996) Theories of theories of mind. Cambridge University Press, Cambridge
Currie G (1998) Pretence, pretending and metarepresenting. Mind Lang 13(1):35–55
Currie G, Ravenscroft I (2002) Recreative minds. Clarendon Press, Oxford
Davies M, Stone T (eds) (1995a) Folk psychology. Blackwell, Oxford
Davies M, Stone T (eds) (1995b) Mental simulation. Blackwell, Oxford
Dennett D (1978) Beliefs about beliefs. Behav Brain Sci 1:568–570
Dokic J, Proust J (eds) (2002) Simulation and knowledge of action. Benjamins, Amsterdam
Gallagher S (2005) How the body shapes the mind. Oxford University Press, Oxford
Goldman G (1996). Simulating Minds. Oxford University Press. New York.
Gordon R (1995a) Folk psychology and simulation. In: Davies M, Stone T (eds) Folk psychology. Blackwell, Oxford
Gordon R (1995b) Simulation without introspection or inference from me to you. In: Davies M, Stone T (eds) Mental simulation. Blackwell, Oxford
Gordon R (1996) Radical simulationism. In: Carruthers P, Smith PK (eds) Theories of theories of mind. Cambridge University Press, Cambridge
Heal J (1994) Simulation vs. theory theory: what is the issue? In: Peacocke C (ed) Objectivity, simulation, and the unity of consciousness. Oxford University Press, Oxford, pp 129–144
Heal J (1996) Simulation, theory and content. In: Carruthers P, Smith PK (eds) Theories of theories of mind. Cambridge University Press, Cambridge
Heal J (1998) Co-Cognition of off-line simulation: two ways of understanding the simulation approach. Mind Lang 13(4):477–498
Jacob P (2002) The scope and limits of mental simulation. In: Dokic J, Proust J (eds) Simulation and knowledge of action. Benjamins, Amsterdam, pp 87–109
Kaplan D (1989) Demonstratives. In: Almog J, Perry J, Wettstein H (eds) Themes from Kaplan. Oxford University Press, New York
Nichols S, Stich S (2000) A cognitive theory of pretense. Cognition 74:115–147
Onishi KH, Baillargeon R (2005) Do 15-month-old infants understand false beliefs? Science 308(5719):255–258
Perner J (1991) Understanding the representational mind. The MIT Press, Cambridge
Prior AN (1963) Oratio Obliqua. Reprinted. In: Prior AN (1976) Papers in logic and ethics. Duckworth, London, pp 147–158
Prior AN (2003) Papers on time and tense. Oxford University Press, Oxford
Recanati F (2000) Oratio Obliqua, Oratio Recta. An essay on metarepresentation. MIT Press, Cambridge
Steward H (1997) The ontology of mind. Events, processes, and states. Clarendon Press, Oxford
Surian L, Caldi S, Sperber D (2007) Attribution of beliefs by 13-month-old infants. Psychol Sci 18:580–586
Wimmer H, Perner J (1983) Beliefs about beliefs: representation and the containing function of wrong beliefs in young children's understanding of deception. Cognition 13:103–128

Chapter 22
Numerals and Word Sequences

Roberto Casati

Abstract According to Spelke and Tsivkin numerals are a linguistic and cognitive bridge between two types of 'core' knowledge, that is, subitization of small quantities and approximate representation of large quantities. In this chapter, I go somewhat their way but I also introduce some a priori constraints on what could constitute a bridge. Such constraints are on the 'design' of a numeral system and on its use. The starting point is the consideration that numerals like 'three' (as well as names of days of the week like 'Friday') are nonstandard linguistic items. I propose that their peculiarity is primarily neither a syntactic nor a semantic peculiarity. It is instead in their morphology. Mastering numerals and names for days of the week is assigning them a certain nonstandard morphology, whereby any numeral is mandatorily a non-independent part of a longer sequence. It is hypothesized that this nonstandard morphology is associated with a nonstandard (at least for language) semantics, i.e. map semantics. In a sense, numerals are an artificial language encroached in natural language. The explanatory advantages of the account are discussed and contrasted with Spelke and Tsivkin 'bridge' account of the role of numerals in cognition.

Keywords Numerals · Maps · Number cognition · Public representations

Numerals Is there anything special, from the linguistic point of view, about *numerals* such as 'two' or 'thirty-seven'? On the face of it, they have multifarious, and confusing, syntactic and semantic properties. In 'three is greater than two' numerals appear to function as nouns and to refer to objects. But the objects they purportedly refer to are elusive (Benacerraf 1965). In 'my friends are three', 'I have three friends' or in 'there are three apples on the table' the numeral appears in an adjectival form to express a property of a set (Frege 1883) or of a manifold (Husserl 1891; Simons 1982). However, they are not like standard adjectives (for instance, in many languages in which adjectives can be pluralized, numerals normally are not). Moreover, the property they purportedly express is mysteriously different from ordinary properties, such as the one expressed in 'my friends are nice'. Finally, sets and

R. Casati (✉)
Institut Nicod, CNRS (EHESS-ENS), Paris, France
e-mail: rcasati@gmail.com; casati@ehess.fr

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_22,

manifolds, the purported describees of 'three', are metaphysically very different from each other—here too the semantics of numerals is elusive.[1]

There are specific interesting linguistic properties of numerals. In English there are two main types of numerals: primitive numerals and compositional numerals. Primitive numerals are those in which no trace of another numeral is to be found. They typically span the initial series of numerals up to the first base ('one' to 'ten'), and then include some other bases ('hundred', 'thousand').[2] 'Seven' is no more compositional than 'three', but for reasons that will be clear in a while it is possible to draw a further classification line between small- and large-primitive numerals (Hurford 2001). Accordingly, I shall make a threefold distinction here: primitive versus compositional numerals, and, within the class of primitives, *small* numerals like 'three' and *larger* numerals (that is, larger than small ones) like 'seven'. A peculiarity of primitive numerals is that they appear to share properties with a larger class of expressions that includes names of days of the week, of months of the year, of fingers, of letters of the alphabets and of musical notes.[3] Observe, however, that the analogy only holds for primitive numerals. Names for days of the week, for months, for notes, for letters, are not compositional. You cannot form the expression 'Friday-Tuesday', for instance, to name a day.[4]

These features of numerals appear to be contingent and idiosyncratic, at least in the sense that they do not flow naturally from intrinsic properties of the numbers (however the latter be construed). There is no metaphysical difference between the number 3 and the number 28. Surely, some explanation is required.

In order to account for these interesting properties of numerals, it has been suggested that attention be paid to their role in number cognition or mathematical cognition at large. The relevant aspects of the research on number cognition are as follows. Two 'systems' of 'numerosity representation' are generally alleged. The first system represents small numerosities in an exact way. The second system represents large numerosities in an inexact way. The two systems are domain specific: They are triggered by varieties of perceptual inputs that are mutually exclusive. The first system is triggered by collections of up to three to four items; the second system is triggered by larger collections, but not by small collections. The first system delivers a sharp classification of the input: Three objects are never categorized as

[1] A difference is assumed here between sets (abstract entities) and manifolds (concrete entities).

[2] Some particular problems posed by 'eleven' and 'twenty' will not be addressed here. *Some* of the idiosyncrasies of numeral systems are likely to be independent of the main point of this chapter.

[3] Bühler 1934 (Mulligan 1997, p. 201) was probably the first to remark the analogies between these sequences, which he called 'organizers'. It is interesting to note that names for colours are not on this list, whereas names of musical notes are. It may be that these particular sequences are just an extension of the large class of asymmetric idioms ('black and white', 'up and down', 'Tom, Dick and Harry').

[4] Compositionality in the field of numerals is not unrestricted, however, as one cannot form expressions such as 'twenty-twenty'. However, look carefully at the various possibilities: 'two-twenty' is acceptable if used to express the hour of the day.

two, two objects never as three. The response of the first system is not only accurate, it is very fast; the system is said to 'subitize' cardinality. This means that in order to judge—at the personal level—that there are three objects when we look at the following image: we do not have to *count*—at the personal level—the items. The second system represents cardinalities in an imprecise way. A collection of 53 items simultaneously presented in vision does not give rise to a sharp classification as is shown by the fact that we are at pain in distinguishing such a collection from a similarly presented collection of 54 items. The two systems do not cooperate or mingle: there is neither an approximate representation of small cardinalities, nor a sharp representation of larger than small cardinalities. Finally, the *main* purpose of the first system is to deliver absolute representations ('there are three objects here in front of me'); the *main* purpose of the second system is to deliver relative representations ('this collection is larger than that').

The two modes of operation are documented in adults and, what is more telling, in pre-linguistic infants. The latter findings indicate that the systems appear to constitute a form of 'initial' or innate knowledge which is, of course, language independent, as it is manifested before the onset of language.

A methodological point concerning experiments on infants. Experiments on subitizing may elicit hybrid interpretations: Part of the interpretation of the results could be based on the experimenter's first-personal assessment of the test display, which may be projected into the infant's capabilities. The fact that the infant reliably distinguishes °°° from °° in suitably controlled experiments cannot be taken as evidence that the infant distinguishes twoness from threeness, *without further assumptions* about what exactly gets counted or classified (or even perceived). If we fancy to redescribe the input not as a collection of circles but as a collection of couples of possibly nonadjacent circles, then the infant is distinguishing threeness from oneness: [(°-°)(-°°)(°°-)] versus [(°°)].[5]

22.1 The Role of Numerals: A Tempting Line of Thought

Precision Imprecise access to cardinality was neglected in traditional discussions of the philosophy of mathematics. It is, however, a key ingredient of numerical cognition. Consider collections whose cardinalities differ by one unit. Subjects are good at determining which is the largest collection between °°° and °°°°, and are bad at the same performance involving even slightly larger groups of items (say, 35 and 36 items to be perceptually discriminated.) However, if the respective sizes of the large groups are different by a substantial fragment (a Weber fraction, informally: The size of a just noticeable difference is correlated with the magnitude of the stimulus), subjects are again able to make fairly good approximate assessments

[5] Interference between figural properties and subitization has been documented too (Trick and Pylyshyn 1993).

(Dehaene 1997, for a review of the literature). This means that there is a relatively reliable approximate access to numerosity in comparison to tasks involving large cardinalities. Observe that it is access to numerosity itself, however approximate, that explains the success in comparison tasks; this is why such access is hypothesized. Access to *relative* numerosity, whatever that could be, would not constitute an explanation, but a redescription of the performance.

Why All this Matters for Numerals In order to understand numerals, we could ask where they enter a stage in the development of cognition; what role they could fulfil in view of the data. A tempting line of reasoning is as follows. Consider again the imprecise system, working for large cardinalities. Faced with a collection of 53 items, we would not represent the items as being 53, but as being some indeterminate rather large amount. Our best type of performance would make us detect a difference between that collection and a collection of, say, 43 items. But the comparison would work without us having any sharp idea of the actual cardinality of the terms, provided the difference between the collections is somewhere in the area of the Weber fraction. The tempting line of thought is now this: If we want a sharper, and possibly an absolutely sharp idea of the numerosity of the large collection(s) in front of us all we need is to *count* the items; now counting amounts to mastering a certain symbolic system. Hence, mastering of a certain symbolic system can sharpen our representations, and (possibly) can create sharp representations. The symbol system in question is, of course, the fragment of natural language constituted by numerals.

Spelke and Tsivkin (2001) suggest that empirical evidence supports in part this line of thought. They take the two number systems as part of our initial endowment of knowledge, and they hypothesize that conceptual development occurs when a *cognitive bridge*[6] can be thrown between the world of small cardinalities (the domain of the first system) and the world of larger cardinalities (the domain of the second system), in such a way that *precision* (a definitory feature of the first system) can reach into the preserves of the second system. *Natural language* is for Spelke and Tsivkin the bridging factor. Although the details of the bridging are admittedly 'obscure' (Spelke and Tsivkin 2001, p. 71), it is to be expected that the relevant component of language is the fragment of numerals, and that the relevant features of language are (i) the fact that the language is *not domain specific* and (ii) the fact that language is a *powerful combinatorial system*. Being domain a-specific, language can generate representations that apply to entities in both the domain of small collections and the domain of large collections. And the combinatorial aspect of language is the actual engine behind the possibility of linking the two domains. Now, surely the fragment of numerals exhibits the required features (it is domain a-specific and is compositional.) So, there is reason to think that it constitutes the required bridge.

[6] Number cognition is just one example of Spelke and Tsivkin more general hypothesis that natural language bridges different systems of initial knowledge. Another example concerns the integration of space and colour representations (Hermer and Spelke 1996).

The Problem with 'Seven' Although there is a theoretical (as well as an intuitive) appeal to the hypothesis that the fragment of numerals is the bridge between two types of cardinality cognition, the hypothesis can be challenged. First, the explanation that mentions the combinatorial system is incomplete. 'Seven' is a primitive numeral—it is not the result of any combination. If we had positional numerals in base three, so that our 'four' were pronounced 'one-one', and seven 'two-one', then it would be immediately possible to think of the combinatorial feature as *the* relevant link between small cardinalities and not so small cardinalities. But as far as I can ascertain no natural language has a numeral system in base three, and 'seven' is invariably translated into other languages as a primitive numeral. Hence, the combinatorial feature of language is not necessary: A primitive numeral such as 'seven' can denote sharply a collection that none of the two nonlinguistic systems recognizes with the desired precision.

Put yourself in the shoes of a cognitive designer who is out to find a way to bridge the gap between cognition of small numerosities and cognition of large numerosities. You discover that the resources of the system you are trying to build already include precise knowledge about cardinalities of 1, 2 and 3, in no order. You see that it would be desirable to know with equal precision about larger cardinalities, about which the system has the resources for delivering moderately reliable but imprecise representations. You know that language is flexible, insofar as it is not domain specific and is combinatorial. So, you first throw in names such as 'one', 'two' and 'three'. What next? One can see how come you invented 'twenty-three'. But how come you invented 'seven'?

Second, and more important, it is disputable that precision is the key factor in bridging the gap between the two systems. Precision, in Spelke and Tsivkin's account, figures as a *normative requirement*. As such, it is *not* required for subitized cardinalities *just because* these *are* subitized. And neither is it required for pairwise comparing of all types of large cardinalities, as some types *are* comparable just as they are. Precision is a normative requirement *only* for comparisons of large cardinalities that are very similar to each other. As perceptual cognition does not help in these specific cases, it is just natural that counting comes into play. Hence, precision is only required with counting. It is born with counting and does not affect cardinality representations as such. To put the point another way, the precision requirement *cannot* affect subitized representations, as they already are as precise as it can possibly get. And it *does not* seem to substantially affect larger cardinality representations, as is proven by the type of persistent mistakes people make in the experiments in which comparison of *numerals* under time constraints shows effects similar to those of comparison of perceptually presented cardinalities (Dehaene 1997).

22.2 The Map Structure of Primitive Numeral Sequences

Let us enrich the picture. If learning of small numerals were to be informed by the semantics available to the infant, then it would make no particular sense that small numerals be learned in the sequence 'one-two-three', or in any other sequence, for that matter. As the infant has a representation of threeness available, one could as well teach numerals, that is, introduce them into the language, starting from 'three', and entering 'one' later on. This suggests that reciting the 'one-two-three' rhyme in that particular pattern is an activity that may have little to do with cardinality cognition. However, it seems that the learning sequence is mandatory. No one is taught numerals by being taught countdowns, or any other available order. It follows that learning the sequence of numerals is an activity not only done independently of number cognition; it is also an activity with an independent goal.

More generally, primitive numerals are learned independently of a semantic assignment. You do not even try to introduce 'seven' alone. You spend a lot of time instructing children to learn numerals in the appropriate order (Gallistel 1993; Karmiloff-Smith 1992). This effort is comparable to the one of learning a rhyme. But what is this learning effort about? Why invest times and energies in a rhyme?

On top of that, names such as 'one', 'two', 'three', are not introduced independently of the introduction of many other names, such as 'seven' or 'ten'. Why invest time and energy in a rhyme which includes names for numbers the kid has no 'initial' access to?

So these are the data: Numerals up to 'three' are learnt in a nonrandom way, hence, disregarding the fact that their semantics is available to the learner; and they are learnt as ordered *parts* of a nonrandom sequence of other numerals, disregarding the fact that for these other numerals the learner has no semantics available.

What Numerals are, Part One: Two-Dimensional Morphology The proposal, I would like to make is that 'three' is not like, say, 'dog' or 'nice', in spite of its occurrence in contexts that are syntactically similar to contexts in which 'dog' or 'nice' may appear (such as 'three is larger than two' or 'I have three friends'), in a new and interesting sense. Where is the difference? The main point is that 'dog' is not *part* of any other word which is only composed of elements like 'dog',[7] whereas—this is my contention—'three' is literally *part* of 'one-two-three-four…'. The difference in question is a *morphological difference*.

One figurative way to describe the difference is to say that locally a sentence like 'I got three apples' has two morphological dimensions. It has the horizontal dimension of spoken enunciation, and it also has the unpronounced vertical dimension that inserts 'three' into 'one-two-three-four…'.

[7] Hence 'doggy' and 'doghood' do not qualify as counterexamples. A very strong construal of my claim here is that, actually, 'three' is not a word, and the sequence *S* is.

'Three' is then assigned a particular morphology, schematically represented by (*part of* ('three', 'one-two-three-four-five-six-seven-eight-nine-ten')).

I write down the sequence 'three' is part of in full only once; henceforth it will be shortened as '*S*'.[8] There are two restrictions on what counts as a part of *S*. First, *S* is quantized: 'wo-th', which is a part of 'two-three', is not a part of *S* in this sense—it is not morphologically visible. Second, only atomic parts are used: 'two-three' is not a part in the intended sense.

What Numerals are, Part Two: Primitive Numerals are One Map The second aspect of the thesis I am putting forward is that the phonetic and quantized sequence *S* is a *map* (not in the mathematical and trivial sense of a mapping, but in the more mundane and interesting sense of a road map). *S* works *like* a *map,* has map properties, and being quantized, it works like a map whose spatial resolution is given by the size of the pixel. *S* is the *vehicle,* it is what does the representing. An immediate consequence of the thesis is that the learning effort for numerals is explained in a simple way: It takes time and effort to *construct the map,* that is, to *construct the vehicle* for the representation. This is so because the vehicle is one long phonetic sequence, possibly the longest single phonetic sequence the child must learn at this stage.

Maps Semantics and Map Cognition Assume, you already have a map at your disposal; a mundane map, of the type encountered in everyday life. Learning to use a map consists basically of learning to find one's way in space by *orienting* the map. This capability decomposes into two independent, but interrelated capabilities. It must be understood that the representation vehicle works as a set of designators for locations, where the designation is *structured* in a specific way; and it must be learnt

[8] One may use a symbol such as '~…' to indicate that there is a word, the word replaced for the three dots is a proper part of. 'Three' would then be represented morphologically as '~Three'. A way to express this colloquially is to say that 'three' is syn-morphological.

how to *situate* the map. Situating the map is a matter of establishing that a point or a region in the space the map represents coincides with (that is, is literally identical to) a point or a region on the map. This is the function typically performed by 'you are here' pointers, together with the nonrepresentational fact that the map is physically nailed to the location that does the situating. Understanding the designation structure (Casati and Varzi 1999) requires to consider each region of the map as a designator for a region in the world, and to see that such designators are structured. In particular, and typically, they obey some part-whole and topological structures: For instance, a map region *r* is part of a map region *r'* just in case the referent of *r* is part of the referent of *r'*[9], and whenever a map region *r* is between map regions *r'* and *r''*, the referent of *r* is between the referents of *r'* and *r''*.

Back to numerals, once sequence *S* has been learnt (that is, it has been stabilized in long-term memory), its properties are so similar to those of a map—albeit of a very simplified, unidimensional map—that the use of *S* can be taught by appealing to resources from map cognition. These resources include the understanding (i) that in order to go from the referent of a certain map region to the referent of a distinct map region one may have to pass through the referent of a third, further different map region (say, one cannot jump from the referent of 'five' to the referent of 'seven'; one has to go through the referent of 'six'). The resources also include understanding (ii) that the portion of quantized *S* that includes 'one-two-three-four-five' is on the face of it larger than the portion of *S* that includes 'one-two'.

These two features of understanding make map-oriented learning of *S* support an ordinal and a cardinal construal of numerals, respectively. This fact may constitute an advantage over a more restrictive hypotheses, according to which numerals are born ordinals or cardinals.

'Friday' Notice a parallel with names for the days. They are learned as rhymes, are introduced in an ordered way and wholesale, are non-compositional, and once learned, their sequence (call it *W,* short for 'Monday-Tuesday…') can be literally seen (or figuratively, as one should better say 'heard') to have map properties: they can be situated, one cannot jump from the referent of 'Thursday' to the referent of 'Saturday' without passing through the referent of 'Friday', and finally the span 'Monday-Tuesday-Wednesday' is seen (heard) to be larger than the span 'Monday-Tuesday'. Another parallel holds for names of the letter of the alphabet.

Entering Available Representations By drawing a comparison to maps, so far only the semantic *structural properties* of numerals are assumed to be understood; nothing has been said about the intended semantics. But of course, once the map features of *S* are understood, the available, automatic semantics for small numerals can be entered. However, the learner need not assign *any* particular representation to numerals larger than the small numerals. It is enough that the learner be able to use the map, in the sense of being able to operate on it.

[9] Some singular terms in natural language are referentially structured in a part-whole fashion that matches the morphological composition: The referent of 'The county of Exeter' includes the referent of 'Exeter'; however, this is not the general case, as the referent of 'The capital of Italy' does not include, and is rather included in, the referent of 'Italy').

What the Thesis Is Not; and Some Consequences of the Thesis At this point is important to prevent some possible misinterpretation of the main hypothesis, and to indicate some consequences and predictions.

a. The thesis is not that all elements of map use and cognition are retrieved or activated in full, but only that some, relevant aspects are taken into account. In fact, there are important disanalogies that block a full assimilation with maps, such as the quantized aspect of numerals and names for days. A closer analogy would be with a pixelized map.
b. There is a certain holistic flavour to the idea that numerals are somewhat represented wholesale. However, the thesis is different from the thesis that in order to master the concept of three one must have mastered the concepts of one, of two, or even of seven (in the sense in which it is said that in order to master the concept of a teacher one must muster the concept of a pupil). The latter is a semantic thesis I am not endorsing or even suggesting. The idea is instead that we have a tacit morphological representation (the sequence *S*), not a tacit semantic representation. (Constraints on the 'processability' of *S* may be not relevant: It *does take time* and effort to learn numerals.)

What are the consequences? According to the thesis, numerals and names for the days of the week are small 'artificial languages' encroached into natural language. The thesis is compatible with the claim that written numerals may have predated spoken numerals.[10] For if numerals are not part of language, and are incorporated elements, where do they come from in the first place? The answer may be that they got incorporated into spoken language from written language or from iconic representations. If not written language, it could just have been gestures. In either case, we *learned to pronounce a visual image*. *Primitive numerals are a map that you can pronounce*. This would be an interesting case in which *spoken language* is parasitic on something else—upon non-spoken language or upon other public representational instruments.[11]

The account is also compatible with the fact that some cultures use names of body parts as numerals. In point of fact, body parts are not assigned randomly to numbers. Hence, the sum of all body parts used as numerals can well be used as a map in the proposed sense.

On Compositional Numerals and their Representations According to Spelke, people form an analog cardinality representation (not necessarily consciously accessed) when they listen to words like 'fifty-six'. The representation in question is approximate, which explains why people are faster and more accurate in processing 'is fifty-six larger than forty-three?' than they are in processing 'is fifty-six larger than fifty-eight?' (Dehaene et al. 1999). However, for the actual use of 'fifty-six' (not

[10] Boyer (1991), although this is only a speculation so far as I can ascertain. Another problem concerns the use of numerals in illiterate cultures; one should check where they got their numerals from, for instance, from a nearby culture that in turn incorporated it from written language.

[11] It is possible that spoken language incorporates many other nonlinguistic elements, pieces of the external world that we learned to pronounce. Possibly demonstratives are such an incorporation—pronounceable gestures.

in an experimental setting, not under fast response constraints) the induced analog representation is totally irrelevant.[12] It *must* be irrelevant, *because* it is imprecise. The mental proxy of the verbal (pronounceable) 'fifty-six' is not one-to-one correlated with the verbal 'fifty-six', as it could fit the verbal 'fifty-seven'. Moreover, the verbal 'fifty-six', when it is used, does *not* make 'numerosity cognition' more precise, as it *cannot:* numerosity cognition in non-very-small numbers *is* imprecise.

Compare, for example, how calculations about time spans based on days of the week are relatively clumsy. Our 'intuitions' about the number of days between today Tuesday and Friday in two different weeks are confused[13] and are not improved by our acquaintance with subtracting or adding by using numerals. Acquaintance with numerals simply did not improve our numeric *cognition*. We have instead to assign dates to the days, and then subtract and possibly add by *explicitly using* the public numerals. Possibly, some training could modify this state of affairs; but this would not amount to creating new, stable 'representations'.

'Seven' As a By-Product of Constraints on the Learnability of Maps: A Speculation Primitive but large numerals like 'seven' extend the knowledge of small numerals into the next immediately accessible field. It is precisely because 'seven' is close enough to 'three' that it could be grasped easily. At the same time, why invent 'seven' in the first place? Why are not numerals immediately compositional after 'three', which would be the perfect solution to the design problem evoked by Spelke and Tsivkin, of bridging the gap between two initially available systems of numerosity cognition? The solution may come from the map nature of numerals. A narrow, 'contingent', 'historical' account would have it that as primitive numerals typically span from 'one' to 'ten', possibly to some translations of 'twenty' in other languages, they are likely to be born with the use of fingers, possibly both fingers and toes, for counting. Hence, names were needed for the whole set of fingers (or were possibly derived from antecedent names for fingers.[14]) However, 'seven' could still have been replaced by a compositional numeral created out of the small primitive numerals. A more 'structural' account may address this problem. Maps properties are readable only off items that are sufficiently *rich*. Although the part of *S* including only 'one-two-three' does have all the features that can be ascribed to a map (part-whole structure, and mandatory passage through the intermediate regions), still these may be much too poorly exhibited in the map to elicit learning of them. The 'one-two-three' sequence would be too small. A larger sequence would support the generalizations that are necessary to understand the map structure of the sequence. The invention of 'seven' is vindicated.

[12] An old Wittgensteinian lesson.

[13] This is anecdotal evidence in need of empirical confirmation.

[14] Names of fingers and toes are interesting in themselves as they constitute sequences. Note that pianists have to relearn names for fingers as numerals.

22.3 Conclusions

The account proposed here is that the sequence of primitive numerals is represented morphologically as a single unit[15] and that this large unit is partly interpreted the way a map would be.

Some consequences of the thesis have been hinted at: The learning effort for numerals is explained in a simple way, as it takes time to memorize the long sequence; the account is compatible with known aspects of map cognition; it is compatible with the possibility that written numerals have been introduced before spoken numerals; and with the fact that in some cultures numerals are represented by body parts. Furthermore, the map interpretation supports both ordinal and cardinal construals of numerals, and makes a plausible place to the invention of 'seven'. Finally, the account is general insofar as it also covers other non 'numeric' examples, such as names of days of the week, of months, of fingers, of letters of the alphabet, of musical notes. Numerals are in no way special.

Against Spelke and Tsivkin, numerals are not what makes it possible to improve number cognition. They are *all there is* as the next step in number cognition. Put otherwise, there is no number cognition on top of the first, subitizing step, and there is no number cognition for large numbers in the sense that the approximate system cannot be 'made' precise by the use of numerals. 57 is invisible to cognition. In this deflationary sense the present account resonates with Wittgensteinian views: 'Our language can be seen as an ancient city: A maze of little streets and squares, of old and new houses, and of houses with additions from various periods; and this surrounded by a multitude of new boroughs with straight regular streets and uniform houses' (*Philosophical investigations,* 18) However, the sequence of primitive numerals does not fall in either category. It is neither a clumsy part of the old city, nor a neat part of the suburbs. It is like a fallen deity, a beautiful crystal placed in the very centre of our hamlet.

Acknowledgments Thanks to Luca Bonatti, Valentina Gliozzi, Pierre Jacob, Nirmalangshu Mukherji, Marco Panza, Achille Varzi, for useful comments on earlier versions of this chapter. And of course thanks to Kevin Mulligan, who got so many projects on the right track.

References

Benacerraf P (1965) What numbers could not be. Philos Rev 74:47–73
Boyer CB (1991) A history of mathematics, 2nd edn (revised: Merzbach, UC). Wiley, New York
Bühler K (1934) Sprachtheorie. Fischer, Jena
Casati R, Varzi AC (1999) Parts and places. The MIT Press, Cambridge
Dehaene S (1997) The number sense. Oxford University , Oxford
Dehaene S, Spelke L, Pinel P, Stansecu R, Tsivkin R (1999) Sources of mathematical thinking: behavioral and brain-imaging evidence. Science 284:970–974

[15] Classifying *S* as a *unit* is, of course, relatively loose talk. It is not clear that *S* is a *word*.

Frege G (1883) Grundlagen der Arithmetik, eine logisch-mathematische Untersuchung über den Begriff der Zahl. Nabu Press, Breslau
Gallistel CR (1993) The organization of learning. The MIT Press, Cambridge
Hermer L, Spelke ES (1996) Modularity and development: the case of spatial reorientation. Cognition 61:195–232
Husserl E (1891) Philosophie der Arithmetik. Pfeffer, Halle
Hurford JR (2001) Languages treat 1–4 specially. Mind Lang 16:69–75
Karmiloff-Smith A (1992) Beyond modularity. The MIT Press, Cambridge
Mulligan K (1997) The essence of language: Wittgenstein's builders and Buhler's bricks. Revue de Métaphysique et de Morale 2:193–215
Simons P (1982) Number and manifold. In: Smith B (ed) Parts and moments. Studies in logic and formal ontology. Philosophia Verlag, München, pp. 160–199
Spelke ES, Tsivkin S (2001) Initial knowledge and conceptual change: space and number. In: Bowerman M, Levinson S (eds) Language acquisition and conceptual development. Cambridge University Press, Cambridge, 70–97
Trick LM, Pylyshyn ZW (1993) What enumeration studies can show us about spatial attention: evidence for limited capacity preattentive processing. J Exp Psychol Hum Percept Perform 19(2):331–351

Chapter 23
Frege's New Language

Jonathan Barnes

Abstract Many philosophers have complained about the inadequacies of natural languages. None more so than Frege—who found German so inadequate for his own peculiar purposes that he invented a new language. How did his new language trump German? Unlike German, the new language contained no ambiguities. So what? Well, ambiguities can mislead—but if you are misled, it is your own fault, not the fault of the language. Again, German disguises the logical structure of the thoughts which it expresses: Grammatical structure and logical structure do not always coincide; and as a result you cannot felicitously formulate patterns of inference by means of German matrixes or schemata. But you can always use metalogical descriptions rather than matrixes, and then German sails home. Again, you can express invalid inferences in German: The syntax of Frege's language ensures that any well-formed inference is valid. But you may still misinfer. True, your error will be syntactical rather than logical: So what? There is still the ace of trumps. Frege needed to express some exceedingly complicated thoughts—notably, thoughts which involve multiple quantifications and thoughts which involve embedded conditionals. Some of those thoughts cannot be expressed in German, which—like other natural languages—has an upper bound of complexity. (No German sentence begins with five occurrences of the word 'wenn'). Frege's language has no upper bound. So the new language has a single advantage over German—an advantage which mattered to Frege (and to virtually no one else).

Keywords Ambiguity · Begriffsschrift · Complexity · Frege · Syntax

It was Kevin and the oysters who clinched it, and I left the banks of the Isis for the richer banks of the Rhone. Along with the oysters went talk of many things—not in our case (so far as I recall) of ships and sealing wax, and still less of cabbages, but much (which I do recall, with fondness) of logic and of the great Gottlob. May these pages summon up the remembrance of things past.

Philosophers have always griped about language. Sometimes the complaints have been local, sometimes global. Some critics have wrung their hands, others

J. Barnes (✉)
Université Paris-Sorbonne, les Charmilles, Ceaulmont, France
e-mail: jonathanbarnes@wanadoo.fr

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_23,

have acted. Frege, from the beginning of his career to its end, was a griper—a global griper, and a man of action. He invented a new artificial language which was to be immune to the diseases which debilitate natural languages and make them incapable of doing scientific service. The new language was a *Begriffsschrift,* an ideography. It is usually referred to as 'the *Begriffsschrift*'; but since it is not the only *Begriffsschrift* in the world, I prefer to call it 'Fregean'.

Fregean came in two versions, Mark 2 being an enlargement and a refinement of Mark 1. Frege never claimed that it was a perfect language full stop—I daresay he thought that no language could be perfect full stop. But he held that it was adequate for arithmetic, and for logic—or at least for that part of logic which arithmeticians need; and he believed that it could be augmented so as to become adequate for any scientific purpose.

What is adequacy? Fregean, I suppose, will be adequate for arithmetic and logic insofar as, first, it is capable of expressing any arithmetical or logical thought; secondly, those formulas which express such thoughts express nothing extraneous to the thoughts which they express; and thirdly, the structure of its formulas corresponds to the structure of the thoughts which they express. Fregean should be complete, and unadorned, and perspicuous.

Is Fregean adequate? Is it superior to Frege's native German? And to my native English? Here are a few loosely connected ruminations on some parts or aspects of those questions.

In a scientifically adequate language, no formula expresses more than one thought. In other words, the language avoids ambiguity, at least at the sentential level. English revels in ambiguity. In its refined version, Fregean is entirely free from ambiguity. (In Mark 1 every single sign is ambiguous).

Frege sometimes suggests that ambiguity is bad because it is misleading—or because it does not lead at all. Faced by an ambiguous expression, you may choose the wrong sense—or you may be flummoxed. Or, perhaps worse, you may not realize that the expression is ambiguous. Aristotle, who liked to sniff out Greek ambiguities and was proud of his nose, thought that philosophers had sometimes gone astray in their reasonings because they had not recognized an ambiguity—Parmenides' metaphysics, he says, floundered because its author had not noticed that the Greek font 'ei\nai' (epsilon, iota + circumflex + smooth breathing, nu, alpha, iota) has several senses. Aristotle was mistaken in his diagnosis of Parmenides (and his own account of the senses of 'ei\nai' (epsilon, iota + circumflex + smooth breathing, nu, alpha, iota) is also mistaken); but no doubt some philosophers have sometimes been fooled by sophists who palter with them in a double sense.

How did Aristotle react to the ambiguities he detected in Greek? Did he think of inventing a new language? Not for a moment. Did he decide to remove all ambiguous words from his own Greek idiolect? Not at all. He continued to speak and write in natural Greek; he used, cheerfully, words which he took to be ambiguous; and occasionally—when it appeared useful—he indicated which sense of an ambiguous expression was the pertinent one. On the whole, that works pretty well: when Bernard gets a puncture, he does not buy a new tractor.

However that may be, Frege, unlike Aristotle, was not particularly exercised by lexical ambiguities: What excited him more were functional ambiguity and

structural ambiguity, and to one particular case of functional ambiguity he returned again and again. In English, phrases of the form 'The ϕ' perform a variety of functions. For example, 'The ϕ' may serve to designate an individual object, as in

> The walrus and the carpenter were walking hand in hand.

Or it may have a generalizing sense, as in

> The mouse is a creature of great personal valour.

Or it may denote a group of objects, as in

> The scribes on all the people shove.

And there are other uses aplenty, not all of them marginal or quirky—

> The glass is falling hour by hour.
> You won't hold up the weather.

The English definite article is, as they say in francophonia, polyvalent.

Mark 1 Fregean has no sign which corresponds directly to any of those uses of the definite article. Mark 2 introduces a symbol—a thick backslash—which answers, roughly, to the use of the definite article to designate an individual object. The symbol is not ambiguous: It cannot, for example, be used to express generality.

So in that respect, Fregean is different from English. Is it also superior? What is wrong with the polyvalence of 'The ϕ'? Not, surely, that it misleads or bamboozles: How many English speakers have been puzzled by Lewis Carroll's line? How many have thought that poor Kit Smart was referring to an individual mouse? How many have scratched their heads and wondered to what individual object the expression 'the weather' refers? No doubt a polyvalent expression always has, as it were, the potential to mislead; but the potential is rarely actualized—and it is, in any event, scarcely dangerous enough to require the invention of a new language.

There are other objections to polyvalence—or rather, to the particular polyvalence of 'The ϕ'. The one most pertinent to Frege's concerns might be illustrated like this. From

> The walrus was walking

you may infer

> Something was walking.

The inference is a special case of a general form; and you might think to specify the general form in something like the following way:

> From 'The ϕ Fs' infer 'Something Fs'.

But that will not do—after all, it allows you to infer

> Something is balmy

from

> The weather is balmy.

It is the polyvalence of the definite article which causes the trouble; for if you use polyvalent expressions in the specification of rules of inference, then your rules will let in fallacy.

That would be serious were it true. But it is not true; you can take the sting out of polyvalence by indicating which is the pertinent value. Do not write:

> From 'The ϕ Fs' infer 'Something Fs'.

Try instead something like:

> From 'The ϕ Fs', when 'The ϕ' functions as a singular designator, infer 'Something Fs'.

That rule will pass the good inferences and stop the bad.

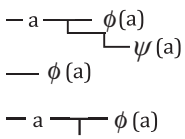
But there is more to the matter. Consider the thoughts you might express in English by the following three sentences:

> The horse eats hay.
> Frege eats hay.
> Nothing eats hay.

The thoughts are different from one another—not just in their truth-value and in their sense but also in their logical structure. But the three English sentences have the same structure. At any rate, had I been asked, half a century and more ago, to parse them, I should have had no difficulty: 'Subject, verb, object, sir'. The structure in question is

> S + V + O.

In Fregean—were Fregean expressively rich enough to talk about hay—the three thoughts would be expressed by three formulas of the following forms:



Those three structures are quite different from one another—and also quite different from the English structure:

> S + V + O.

If a condition on the scientific adequacy of a language is that the structure of its formulas corresponds to the structure of the thoughts which they express, then here is a clear case in which Fregean is adequate and English is not. And the trouble is caused, in part, by the definite article, which disguises the structure of the thought that the horse eats hay.

To that claim, a radical objection is sometimes made. It runs like this. The claim supposes that we can distinguish between, and then compare, two items: the

structure of a given thought, and the structure of a sentence which may be used to express it. But how can we determine what the structure of a given thought is? Well, you cannot inquire into the structure of a thought without identifying it as the thought that *P*; and you cannot do that except by producing some sentence, *S,* which expresses the thought that *P*. In that case, how could the structure of the thought that *P* fail to match the structure of the sentence *S*? To be sure, the thought that *P* might be expressed equally well by two sentences, *S* and *S*,* which have different structures. But that is not a difficulty: It will follow that the thought that *P* has (at least) two different structures; and that is scarcely news.

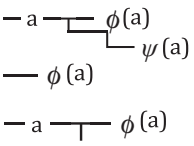
In that case, no language is structurally more apt than another. If the thought that the horse eats hay is expressed in English by the sentence

The horse eats hay,

and if that sentence has the structure:

S + V + O,

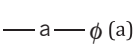
then the thought has the noetic counterpart of that structure. No doubt the thought also has a structure which is the noetic counterpart of the Fregean:



But that does not put Fregean ahead of English. The two languages are equal on points.

What is 'noetic structure'? There is the closest connection between the structure of a thought and its inferential powers: roughly speaking, the structure of a thought—better: the structures of a thought—determine the ways in which the thought performs in formal inferences. Thus, the thought that Frege eats hay is entailed by the thought that everything eats hay, and it entails the thought that something eats hay. Those inferential powers are fixed by the noetic structure—or rather, the expression 'noetic structure' (or 'logical structure', or 'structure' tout court) is, so far as I can see, nothing but a metaphorical manner of speaking about inferential powers.

Now, the two inferential powers of the thought that Frege eats hay are shown up by the structures of the Fregean formulas which express the thought. The first entailment is an instance of the formal inference from:



to

The second is a case of the inference from:

$\vdash \phi(a)$

to

$\vdash \exists a \vdash \phi(a)$

And on that point, Fregean is superior to English. For there is no formal inference from:

S + V + O

to

'Something' + V + O.

For example, from

Nothing eats hay

it does not follow that

Something eats hay.

So, Fregean is ahead on points.

But English need not throw in the towel. No doubt, Fregean gets its structures right in the way I have just sketched. But English is not limited to the schoolboy structure:

S + V + O.

For example, there is a familiar distinction between 'surface' grammar and 'deep' grammar. The surface grammar of the three English sentences under discussion may indeed be characterized by the schoolboy structure; but there is also the deep grammar—which will doubtless turn out very similar to the surface grammar of the Fregean formulas. Perhaps that is true—but it only serves to underline the superiority of Fregean over English. For Fregean, which has no deep grammar, wears on its surface what English conceals in the depths.

Hold on to the towel. Each of the three English sentences has the structure:

S + V + O;

but that is only one part of their surface grammar. The most elementary of English grammar books will point out that there are different sorts of subjects (and different sorts of verbs and of objects, come to that). A subject may be a proper name, or a pronoun, or a phrase of the form 'The ϕ', or…. So although the three sentences share at least one common structure, they also have structures of its own. For example, a slightly more refined grammatical analysis might come up with the following three structures:

PN + VF
[art + CN] + VF
Pron + VF

('VF' is 'verbal formula'). The inference from

Frege eats hay

to

Something eats hay

is an instance of the inference from

PN + VF,
to
'Something' + VF.

The underlying rule does not warrant the inference from

Nothing eats hay

to

Something eats hay.

And English has not yet received the K.O.

But the rule in question does not warrant the inference from

The walrus was walking

to

Something was walking.

For, 'The walrus' is not a proper name. Rather, it is an inference from:

[art + CN] + VF
to
'Something' + VF.

And that form of inference is not generally valid.

Earlier, I said that you might express a rule of inference in the following fashion:

From 'The ϕ Fs', when 'The ϕ' functions as a singular designator, infer 'Something Fs'.

And that rule warrants the walrus inference. It is objected that the rule goes against a principle implicit in contemporary logic, a principle according to which rules of inference should be stated by way of matrixes or schemata. That is to say, rules have the general form:

From a set of thoughts of the forms F_1, F_2 … F_n infer a thought of the form F*.

The forms are expressed by matrixes or schemata—by sequences of dummy letters and significant symbols such that the appropriate replacement of the dummies by significant symbols yields the expression of a thought. The rule I have just commended does not fit that plan; for it does not specify the form of the premiss by means of a matrix. And that is no accident: On the contrary, the rule succeeds—if it succeeds—precisely because it uses more than matrixes to specify forms.

But why should a rule fit the plan? After all, you might think that the rule which licenses the inference of

> Something is walking

from

> The walrus is walking

ought also to license the inference of

> Something eats hay

from

> Frege eats hay.

Those are two particular instances of one general inference; and that general inference cannot be explained by way of matrixes. Rather, you might opt for something like this:

> From a thought which ascribes something to an individual, infer a thought which ascribes the same thing to something or other.

Rules which do not use matrixes will not be muddled by the vagaries of English grammar; for they are neutral between English and Fregean (and between any two languages you like).

And there is something else to be said for them. From the disjunction

> Either he's dead or my watch has stopped

together with

> My watch hasn't stopped

it follows that

> He's dead.

What general rule underlies that inference? Consider this rule:

> From 'Either P and Q' and 'Not Q' infer 'P'.

Very good—except that we shall then need a different rule for the inference from

> Either he's dead or my watch has stopped

together with

> He's not dead

to

My watch has stopped.

Without matrixes a single rule does the job:

From a disjunction and the negation of one of the disjuncts infer the other disjunct.

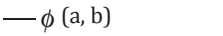
That rule deals with two-membered disjunctions. If you want to state a comparable rule for three-membered disjunctions, or a general rule for *n*-membered disjunctions, then you will get nowhere at all with matrixes.

If the rules of inference are expressed not by schemata but by what the Greeks called periocaiv (π, ε, ρ, o, χ, α, $\acute{\imath}$) or metalogical descriptions, then English is still up and fighting. But of course Fregean can take the same line.

Return to the thought that Frege eats hay. From it you may infer that something eats hay—and also that Frege eats something. How might the inference be dealt with in Fregean? Since eating is a relational affair, you might be briefly tempted by the notion that

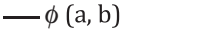
Frege eats hay—so Frege eats something

is an instance of the inference from:



to

But that will not do. A formula of the form:



is well formed only if the symbols which replace 'a' and 'b' are singular designating terms—proper names, in Frege's jargon. But 'hay' is not a proper name: It is a common noun (if you are half tempted by the idea that 'hay' is a proper name—a name for the totality of hay in the universe, say—then take:

Russell eats apples.

That raises exactly the same questions—and there is no temptation to construe 'apples' as a proper name).

Frege recognizes function-expressions which take argument-expressions of different orders in different places: Thus, there are two-placed function-expressions of the form '$\phi(\xi, \mu)$', where the first place is to be occupied by a proper name and the second by a one-place (first-order) predicate. Why not think that 'eat' is such an expression?—that 'eat(ξ, μ)' takes a proper name in its first place and a one-place predicate in its second place? Or, equivalently, that it takes a one-place predicate

in its second place and makes a one-place predicate (that is to say, insert 'hay(ξ)' in the second place of 'eat(ξ, μ)' and—after a bit of idiomatic polishing—you get 'eats hay(ξ)').

That is, I think, good Fregean. But is it as good as English? English has a single relational expression, '… eat —' which appears both in

> Frege is eating hay

and also in

> Frege is eating that *Bratwurst*.

One and the same expression, with one and the same sense, has two different constructions: its second place may be filled either by a common noun or by a proper name. Fregean does not admit expressions of that sort; for in Fregean no expression may have more than one syntactical construction. So, Fregean—if it is to express the theorems of the science of gastronomy—will be obliged to have two predicates, 'eat(ξ, ζ)' and 'eat*(ξ, μ)'. That is hardly a catastrophe. But I incline to think that, in this respect, English is superior to Fregean.

Structure matters in a scientifically adequate language because structure is allied to inference. And an adequate language must be able not only to express inferences perspicuously but also to signal them clearly and distinctly. English, like other natural languages, has various signals, the most common of which are deictic adverbs: 'therefore', 'so', '*ergo*'…. Frege claimed that these natural signals are unsatisfactory, and for two reasons.

First, he noticed that the inferential adverbs of natural languages are promiscuous—that they will associate with any inference which smiles upon them. The word 'therefore' may introduce the conclusion of an induction or of an enthymeme or of a deduction, and when it introduces a deductive conclusion, that conclusion may follow (or be supposed to follow) in accordance with any of an indeterminate number of rules. In short, the inferential adverbs serve to introduce a conclusion but do not specify how the conclusion is supposed to be inferred.

That is true. But why complain? As well, object to the adverb 'later' because it indicates that one thing happens after another but does not specify how long after. That is an advantage rather than a disadvantage. If you cannot or do not want to specify how much later something happened, then 'later' is just what you need. And if you do want to specify, add a qualifier: 'Ten minutes later…', 'Half a lifetime later…'. Similarly with the inferential adverbs. They are just the ticket when you cannot or do not want to indicate what sort of inference you are drawing; and they are readily reinforced when you do—'Therefore, by a syllogism in Barbara…, 'So, by induction…'.

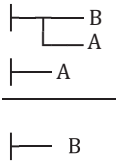
The second of Frege's complaints about 'therefore' is this: in English, good grammar does not guarantee good logic. Here is an inference I made the other day:

> The second letter of 'skate' is k—so a kea must be a New Zealand bird.

Was it a good inference? You cannot tell without more information. But you can tell, without any enquiry, that it was impeccably expressed. As far as grammar is concerned, you may write 'Therefore' or 'So' between any indicative sentences you

like: The result will be grammatically impeccable, even though it may be logically lunatic.

Frege thought that that was a lamentable state of affairs; and in Fregean, good grammar guarantees good logic. In principle, Fregean can express only one sort of inference, namely a version of *modus ponens*. The inferences are expressed by formulas of this form:



The horizontal line which separates the formulas corresponds to the English phrase 'Therefore, by *modus ponens*'.

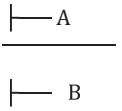
Frege says little about this symbol: He spends a few pages in explaining how expressions for inferences may be abbreviated, and in doing so he makes various modifications to the horizontal line; but the line itself is introduced without remark. Nonetheless, it is a genuine symbol of Fregean—it is not just a punctuation mark. And it may be explained like this: (i) the horizontal line takes three judgement-expressions to make an argument-expression, two of the judgement-expressions being written above the line and one below it; (ii) the upper two judgement-expressions have the forms:

and

and the lower one has the form

and (iii) the line signifies that the judgement expressed below it follows, by the rule of *modus ponens,* from the two judgements expressed above it.

That explanation has the result which Frege wanted: the grammar of Fregean, unlike the grammar of English, guarantees validity—in Fregean, you cannot set down an argument which is not valid. If you inscribe something of the form:



you have not expressed an invalid, or a dubious inference. You have expressed nothing at all—just as you would express nothing at all if you wrote, say:

The grammar of the horizontal line ensures that whenever it appears on the page, as part of a grammatical construction, then it marks the presence of a valid argument in *modus ponens*. There is no means, in Fregean, of marking any other sort of inference; and there is no means of expressing an invalid inference of any sort.

In that way, Fregean differs from English. Is it also, in that way, superior to English? It is true that there are some things you cannot do in Fregean which you can do in English: you cannot propound lousy arguments. But who wants to propound lousy arguments? (You may want to say of a given argument that it is invalid. But in order to do so, you need not propound the argument—it is enough if you can describe or designate it). And is not it a great advantage not to be able to express bad inferences?

To be sure, writing in Fregean cannot stop you from misreasoning. If you write in good Fregean you will never write down a bad argument: it does not follow that you will never write down anything wrong. Fregean does not protect you from error—it merely ensures that your errors will be syntactical rather than logical. Still, it is in fact easier to avoid errors if you write in Fregean than if you write in English. Not because Fregean grammar guarantees validity (though it does do so), but because Fregean gives you fewer opportunities to go wrong, whether you are trying to produce an inference or to check an inference which has already been produced. Suppose you look at an inference on a page of the *Grundgesetze* and wonder whether or not it is valid. Fregean can only express one sort of inference; so you need only ask whether what is in front of you is an inference in *modus ponens* or not. Fregean has only one way of expressing such inferences, so you need only ask whether the symbols in front of you form such an expression or not. And the way of expressing the inferences in Fregean makes it easy to tell whether or not the symbols form expressions of the right sort.

True, in real life it is not quite so simple. First, the inferential abbreviations which Frege introduces in order to save ink make it harder to check an argument for validity: try reading the last section of *Begriffsschrift* or pretty well any section of *Grundgesetze*. Secondly, one of the things you need to check is that ‘A’ and ‘B’ are each replaced twice by the same expression, and since the expressions which replace ‘A’ and ‘B’ may have any degree of complexity, it is not difficult to overlook a minor difference. Nonetheless, the substantial point remains: arguments in Fregean are easier to check than the corresponding arguments in English—and that sort of perspicuity is an advantage.

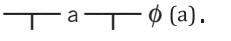
Still, what gives Fregean the edge is the fact that it limits himself to a single style of inference and a single way of expressing it. Suppose, you want to stick to English

but would like to enjoy the advantages of Fregean: is not it enough to state clearly, at the start, that the only form of inference which you are going to use is *modus ponens,* and that the only way in which you are going to express such inferences is by sequences of the form:

> If P, then Q.
> P.
> Therefore Q.

A reader who wants to check your arguments is as well off as if you had written in Fregean: He need only ask 'Is this in *modus ponens*?', and to answer he need only check whether or not it has the prescribed form.

Consider this Fregean formula:



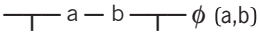
A 'literal' English translation might be:

> Not everything doesn't ϕ.

That lumbers—and an English speaker will prefer the simple equivalent

> Something fs.

Now make the predicate two-placed:



That may be Englished by:

> Not everything doesn't ϕ everything,

which is intelligible if heavy. But there are, of course, predicates with more places—for example, the predicate you arrive at by abstracting the four proper names from the following sentence:

> Camilla gave Archie to Marie at Christmas.

(Archie is a cat.) From that predicate, together with quantifiers and negations, you can arrive at a large (and calculable) number of items, each of which can be clearly and distinctly expressed in Fregean. Expressing them in English at all is difficult, and expressing them clearly and distinctly is next to impossible. What will an English speaker make of, say:

> Nobody ever gave anyone everything?

Here is the moral: certain complex thoughts can be expressed easily in Fregean, painfully in English.

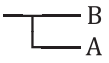
Consider next a few sentential operators: ‘It’s false that…’, ‘No-one believes that…’, ‘England expects that…’—such things can be iterated, and embedded, and mutually permuted. Up to a point the complexity is fun: ‘…at least, I knew he thought I thought he thought I slept’. But sooner or later you arrive at the limits of intelligibility. Those particular operators are, it is true, of no professional interest to Frege. But Frege was essentially concerned with the operator which produces conditional sentences. Fregean expresses conditional thoughts by way of the connector:

The rule for well-formedness is this:

If ‘A’ and ‘B’ are well-formed sentences, then

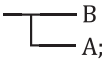
is a well-formed sentence.

The rule guarantees that Fregean contains an unlimited number of variously complicated sentences of the form:



For ‘A’ and ‘B’ may, either or both, be replaced by formulas of the form:

and in those replacements, ‘A’ and ‘B’ may, either or both, have the form:



and so on without end.

One ordinary way of expressing conditional thoughts in English uses the connector ‘If…then—’. It sounds plausible to say that

> If '*P*' and '*Q*' are well-formed indicative sentences, then 'If *P*, then *Q*' is a well-formed indicative sentence.

If that is so, then English, like Fregean and for the same reason, contains an unlimited number of well-formed sentences of the form

> If *P*, then *Q*.

So, in this respect, Fregean and English are on a level.

Is the rule for English correct? If it is, then there are English sentences which begin with five, or fifty, or five hundred occurrences of the word 'if'. Two ifs are ungainly but manageable—for example:

> If if Australia beat England then Australia play South Africa, then South Africa is in the final.

Three ifs are harder. I once dredged up something semi-intelligible with four ifs. Five defeat me. As for fifty.

So if the rule is correct, there is an infinite number of well-formed English sentences which no English speaker could possibly understand. That is surely absurd: How could there be English sentences which no master of English could comprehend? The rule needs to be replaced by something like this:

> If '*P*' and '*Q*' are well-formed, and neither goes beyond a certain degree of complexity, then 'If *P*, then *Q*' is well-formed.

What degree of complexity? Presumably, there is no sharp boundary between the well-formed and the ill-formed, no commandment of the sort: Thou shalt have but three ifs. Rather, well-formedness shades into not-so-well-formedness, and that shades into ill-formedness.

When sentences become complex, English differs from Fregean. And there are some thoughts—an infinite number of them—which can be expressed in Fregean and not in English. Or are there? What goes for English goes for all natural languages. Why does not it go for Fregean too? Well, Fregean is an artificial language, and the rules for well-formedness are determined by its artificer. Frege specified the rule of well-formedness for his conditional stroke, and it follows that Fregean contains sentences which deploy five or fifty or five hundred such connectors. English is a natural language: its rules are discovered, not invented.

You might allow that there is a formal or theoretical difference between English and Fregean and deny that there is any substantial or practical difference. After all, Fregean has its own limits of intelligibility, and there are well-formed sentences in Fregean which no master of Fregean could comprehend. When the limits of intelligibility are crossed, English comes to a stop: Fregean marches on, for ever. But what on earth could be the advantage of that? What could be the point or purpose of generating sentences which no-one can understand? (Or perhaps theologians should write in Fregean?).

But in fact there is a substantial and practical difference. To be sure, it is a contingent difference—but no less significant for that. The difference is this: There are some scientific thoughts which can be expressed and understood in Fregean but

cannot be expressed in English. To see that that is so, it is enough to turn the pages of the *Grundgesetze*. There you will find any number of complicated formulas written in Fregean. The complicated formulas are not easy to understand, not even once you have familiarized yourself with Fregean. But they can be understood: their sense can be worked out, even when it cannot be taken in at a glance. There are no English translations of the formulas.

English is inadequate inasmuch as it cannot express all the thoughts which Frege needs to express: Fregean can. And it may be added that, in this respect, Fregean is superior not only to English and other natural languages but also to the artificial languages which logicians have universally preferred to it. Translate the *Grundgesetze* into Peano-Russell, for example, and you are quickly lost: formulas stretch over two or three lines, and the brackets and dots which punctuate them are dizzying. As Frege realized, the superiority of his language derives in part from the fact that it is two-dimensional—natural languages are all one-dimensional, and so are the artificial languages of contemporary logic.

Fregean is superior to other languages primarily and essentially because it can express in an intelligible fashion thoughts which they cannot express intelligibly or cannot express at all. That, I suppose, is an unremarkable conclusion to arrive at. After all, Frege invented Fregean because the further he advanced in his project, the more tortuous and the less comprehensible became his German; and he invented Fregean because he needed a language which could express intelligibly certain sorts of highly complex thoughts. Fregean has the virtue it was designed to have.

'If Fregean is so bloody marvellous, why has no-one but Frege ever written in it?' (Well, Carnap once wrote a postcard in Fregean—but that does not count). The answer is simple: What distinguishes Fregean from other languages is its capacity to express extremely complicated thoughts. You have little reason to learn Fregean unless you desire to express thoughts of that kind—and that's the rarest of desires.

Chapter 24
On Liars, 'Liars' and Harmless Self-Reference

Wolfgang Künne

Abstract The topics of this chapter are (1) the history of a mislabelled antinomy and of a pseudo-paradox and (2) some logico-semantical peculiarities of self-referential sentences that do *not* give rise to a paradox. My points of departure will be Bernard Bolzano's discussions of a plain fallacy *he* called The Liar and of an antinomy that *we* unfortunately got used to calling The Liar. He found a pointer to the fallacy in Aristotle's *Sophistical Refutations*. In a logic manual of the early renaissance, he came across a source of the antinomy in the form of a sentence that declares itself to be false. In Sect. 24.1, I shall praise Bolzano's reaction to the fallacy and discuss his analysis of the concept of lying. I will present some ancient expositions of the antinomy and go on to criticize, along Moorean lines, Russell's rather sloppy account. Finally, I will defend the author of the 'Letter to Titus' against the charge of being paradox-blind when he invoked a Cretan denigrator of all Cretans. (Some twentieth century logicians and analytic philosophers are the villains of this part of my chapter: I shall criticize their carelessness with respect to a well-entrenched concept, and I shall complain that they keep on alluding to ancient texts without bothering to read them closely.) In Sect. 24.2, I shall reconstruct Girolamo Savonarola's excellent exposition of the antinomy, examine Bolzano's criticism of the Florentine diagnosis and reject his own attempt to defuse the paradox. (I shall not try to improve on his attempt.) In this context, Bolzano makes a point concerning self-referential sentences that is not affected by the failure of his alleged dissolution of the antinomy. He rightly takes it to be a matter of course that there are ever so many harmlessly self-referential sentences. But he shows that some care is needed when one wants to formulate their *negation*. In Sect. 24.3, I will expound this point. It turns out that similar problems arise when one uses harmlessly self-referential sentences in *deductive arguments.* Such sentences also enforce a revision of certain intuitively plausible constraints on *translation*.

Keywords Epimenides · Liar · Lie · Self-reference · Translation

---

To Kevin Mulligan—*with affection.*

W. Künne (✉)
University of Hamburg, Hamburg, Germany
e-mail: wolfgang.kuenne@uni-hamburg.de

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_24,

24.1 Liars, 'The Liar' and The 'Epimenides'

24.1.1 The Fallacy of the Self-Confessed Liar

In Bolzano's own subject index to his monumental *Wissenschaftslehre* (1837) you find the entry '*ψευδόμενος* (pseudómenos), *Trugschluß*' (Bolzano, *WL IV* 673; original pagination[1]). He leaves the Greek word untranslated (and for a while I shall do the same, for reasons that will become clear in Sect. 24.1.3), and he classifies the bearer of the Greek nickname as a *fallacy*. The long chapter that is referred to (*WL* III, § 377) is indeed devoted to the analysis of a couple of famous fallacies, and the structure of this chapter is modelled on an ancient paradigm:

> Aristotle who wrote a special book on this subject (περὶ [τῶν] σοφιστικῶν ἐλέγχων),[2] classifies all false proofs or…fallacies under two genera: those in which the deception arises from the verbal expression chosen…and others, where this is not the case. This classification is quite correct; only one should not forget, as Aristotle himself remarks [Aristotle, *SE* 5: 167^{a}35 and *SE* 33: 182^{b}10–15], that one and the same fallacy may belong in both genera. (WL III 479)

In Chap. 25 of his *Sophistical Refutations,* Aristotle examines a species of fallacies that he takes to belong to the second genus. Its Latin name is *fallacia secundum quid et simpliciter.* Such a fallacy occurs if in an argument, 'something that is said with certain qualifications (*ἐν μέρει λεγόμενον*)' is understood as if it were 'said without qualifications (*ὡς ἁπλῶς εἰρημένον*)' (*SE* 5: 166^{b}37–167^{a}20, here 166^{b}38–167^{a}1). One could call such arguments 'fallacies of neglecting a necessary qualification'. One commits such a fallacy if one concludes from certain premisses that something is both *F* and not-*F,* while ignoring (or deliberately concealing) the fact that from those premisses, if properly understood, it only follows that something is *F* 'in a certain respect (*πῇ*, *secundum quid*)' and not *F* 'without qualifications (*ἁπλῶς, simpliciter*)', or vice versa. Here are two of the questions to which this distinction is applied in Chap. 25:

> [a] Is it possible for the same man at the same time to be an oath keeper and an oath breaker (ἅμα εὐορκεῖν καὶ ἐπιορκεῖν)? Can the same man at the same time both obey and disobey the same man (ἅμα 1/4 πείθεσθαι καὶ ἀπειθεῖν)? (SE 180a34–36[3])

Aristotle replies:

> [b] If a man is in a given case or in a certain respect an oath keeper he need not be simpliciter an oath keeper. (If somebody has sworn on oath to break an oath and finally does break

[1] Abbreviations are explained in the bibliography.

[2] This opusculum, presumably written around 355 BC, is a kind of appendix to the topics. Already in WL I 13 Bolzano refers to it—with the somewhat idiosyncratic title 'Die Sophistik'.

[3] Translations from Greek, Latin and German are always mine. There is already a small problem with [**a**] Julius Hermann v. Kirchmann, SE-(A) [see Bibl.], and Eugen Rolfes, SE-(B), render the verbs in the second question, 180a36, and in [**c**], 180b1, translated above as 'obey' and 'not obey', by 'glauben' and 'nicht glauben'. But in the Greek of Aristotle's time, though '' can indeed be rendered by believe sb. as well as by obey sb., for 'ἀπειθεῖν' only the use in the sense of 'be disobedient' is testified: see LSJ. The SE-translations (*C*) and (*D*) take this into account.

> it, then he acts in accordance with that particular oath but he is not [sc. simpliciter] an oath keeper.) [c] Similarly, the disobedient man is not [sc. simpliciter] obedient just because he obeyed a particular command. (SE 180a38–180 b2).

Suppose that in spring, Alcibiades takes an oath to break his earlier oath never to betray Athen's military plans, and that in summer he does betray them to the Spartans. Is he an oath keeper in summer? *Secundum quid* he is, Aristotle would answer, but not *simpliciter*. Similarly, a Scythian slave who complies with a particular command of his Greek master might not be *simpliciter* obedient, though in that situation he is.[4] Aristotle goes on to present a third question to which this distinction can be applied:

> [d] The argument is similar when we turn to the question whether it is possible for the same man at the same time to say something that is false and to say something that is true (ψεύδεσθαι ἅμα καὶ ἀληθεύειν). [e] But since it is not easy to see whether we have a case of simpliciter saying something true or a case of simpliciter saying something false, here the issue seems to be difficult (δύσκολον φαίνεται). [f] However, there is nothing to prevent somebody from being simpliciter mendacious (ψευδής) but truthful (ἀληθής) in a particular respect or relation and from there being truth in what he says though he himself is not truthful. (SE 180b2–7.[5])

Unfortunately, Aristotle does not give any examples in [**d**]–[**e**]. After his *SE* had become known in the Latin West around 1130, his commentators tried to fill the gap. Some offered examples like 'I am about to say something that is false. Donkeys can fly.'[6] We should not accept this offer, for it neglects the adverb '*ἅμα* (at the same time)' in [**d**] that echoes an earlier occurrence in [**a**]. In the flying-donkey example we have two statements of which the first is made true by the second, and here the

---

[4] As regards [**b**], Catarina Dutilh Novaes and Stephen Read (2008) 179 conjecture: 'it is not implausible that at least some medieval authors read this passage under [a different] reading, namely that the oath made is: "I promise to break this very oath".' I think I can offer evidence for this conjecture: Albert of Cologne alias Albertus Magnus (*floruit* 1240) seems to have interpreted the passage in that way in his commentary on SE (lib. II, tract. iii, cap. 3, transl. in Bocheński 276). But this reading of [**b**] is very far-fetched: There is no hint of self-referentiality in our text. In Dutilh and Read, the interpretation seems to be due to an urgent desire to find in [**b**] an example that is analogous to 'typical Liar sentences' (loc. cit.). Paolo Crivelli (2004, p. 142, 144) thinks he knows which particular command Aristotle had in mind when he wrote [**c**]: 'Be disobedient!' To be sure, that would be a very interesting example, but again I cannot see that Aristotle could reasonably expect his readers to think of this 'self-destructive' command on the basis of what he says. Scholars who think that [**d**]–[**f**] contains a discussion of the so-called liar paradox (Sect. 1.3) are prone—if I may try to import a German catchphrase into English—'to hear the grass grow' already in the preceding sentences.

[5] There is a serious translation problem in [d]–[e]. The verbs '*ψεύδεσθαι*'/'*ἀληθεύειν*', rendered in *SE*-(*BCD*) and above as 'to say sth. that is false/true', are translated in *SE*-(*A*) by '*lügen* (to lie)' and '*die Wahrheit sagen* (to speak the truth)'. So Kirchmann takes *sincerity* to be the point at issue. This divergence will soon play a major role in this chapter.

[6] Egidio Romano [fl. 1290], Supra libros elenchorum Aristotelis, quoted in Paul Spade (1973, II. pp. 305–306) where the author is called Giles of Rome. A commentary on SE that is a century older contains an example of the same sort: Spade (1973, p. 303). Spade also criticizes these attempts to exemplify Aristotle's point. Apropos 'floruit': In order to save space I use 'fl.' even if the birth and death dates of a philosopher are known, and in order to determine the pertinent year I have made an assumption that puts me into a somewhat melancholy mood: Philosophers flourish at the age forty.

distinction between *simpliciter* and *secundum quid* is entirely irrelevant. We need examples where one and the same statement seems to be a contender for both truth values and where this appearance can be dispelled by applying the Aristotelian distinction. Here are some examples that fit the bill. If somebody were to say 'Watermelons are red', his statement would be *simpliciter* false, but because of the pulp of watermelons it would be true *secundum quid*. If Meletus asserts 'The man over there looks like Socrates' while pointing at a snub-nosed man who is arguing with Euthyphro, his assertion is *simpliciter* true but it is false *secundum quid* because it suggests a falsehood, namely that the man referred to is not identical with Socrates.

While [**d**] and [**e**] are about a man who *says* something true or false, [**f**] is about somebody who *is* 'true' or 'false'.[7] We no longer use these adjectives in the way Aristotle used their Greek counterparts (so I avoided them in my translation), but Shakespeare did: Cleopatra asks Antony, 'Why should I think you can be mine, and true, who have been false to Fulvia?'[8] Aristotle concedes that the cases characterized in [**d**] and [**e**] are not unproblematic. But he takes it to be a matter of course that the situation described in [**f**] is possible: Occasionally persons who are mendacious suffer from a fit of veracity.[9]

Bolzano explicitly mentions Aristotle's division of fallacies before he turns to the argument of the Self-Confessed Liar (as I propose to call it). I surmise that in his exposition and evaluation of that argument he took a clue from passage [**f**] in Aristotle's text:[10]

---

[7] The contrast is marked by the elided 'de' that I rendered as 'however'. This transition is correctly rendered in the German translations (*A*) und (*B*) and in Jonathan Barnes' (D), whereas William A. Pickard-Cambridge made it invisible in (*C*): 'There is, however, nothing to prevent it [!?] from being false absolutely, though true in some particular respect….' Spade (1973, II. p. 301) relies on (*C*) although one page earlier he had quoted a Latin translation that avoids this mistake. Peter Eldridge-Smith (2004, p. 78) still quotes (*C*). In Bocheński (152) the passage is mistranslated in the same way as in (*C*).

[8] *Antony and Cleopatra* (I/3: 27, 29). The German counterpart was similarly used by Friedrich Schiller: '*O wärst du wahr gewesen…alles stünde anders!*" Max Piccolomini says to his father, '*Unselge Falschheit…Du jammerbringende, verderbest uns! Wahrhaftigkeit, die reine, hätt' uns alle…gerettet*' (*Wallensteins Tod*, II/7). And Philipp II says about his wife: '*So ists erwiesen, sie ist falsch*' (*Don Carlos*, III/1).

[9] Dutilh and Read (2008) overlook this matter-of-course feature of [**f**]. They also rely on *SE*-(*C*): 'Aristotle adds (sc. in $188^{b}5–7$ = [**f**]) that it is possible for something (a *proposition*) to be false absolutely, though true in some particular respect…It is a telling fact that no mention is made of the converse…We shall see that [Thomas] Bradwardine [in his work on logical paradoxes] will indeed treat the latter case as impossible' (op. cit., 179/180, my italics; cp. 177 n.).' Whatever the value of that treatment by the Oxford mathematician and theologian [*fl.* 1330] may be, it is certainly *not* a development of 'Aristotle's original idea' (184), i.e. of 'Aristotle's undeveloped hunch [that] it is possible for a proposition to be true *secundum quid* while false *simpliciter* (but not the other way round)' (191). The 'idea' or 'hunch' that Aristotle allegedly formulates in $188^{b}5–7$ is only to be found in a bad translation of that sentence.

[10] Jan Berg (1992, p. 109 n. 91) already referred to SE 25: 180b2–7 as a potential source. In his excellent monograph Alexander Rüstow (Appendix II) is mainly concerned with the early history of the so-called Liar Paradox, but he mentions in passing the pages in WL I that I shall deal with in Sects. 24.2 and 24.3 of this chapter (Rüstow 121–122), but he missed the passage in WL III, as did all other scholars whose work I consulted.

1. 'It is possible for a liar to admit that he is a liar.'
2. 'If a liar admits that he is a liar, he speaks the truth.'
3. Thus it is possible that a liar speaks the truth.
4. 'But somebody who speaks the truth is no liar.'
5. 'Thus it is possible that a liar isn't a liar.' (*WL III* 487–88)[11]

In his diagnosis of this argument Bolzano uses the phrase '*die Wahrheit verläugnen* (to deny/to disown the truth)' as antonym of '*die Wahrheit sprechen* (to speak the truth)'. If in an assertion you 'deny the truth' (as Peter did when he said, 'I do not know the man'—*Matth* 26: 69–75) you assert something that is false and that you also take to be false. The same holds mutatis mutandis of the predicate that occurs in (2), (3) and (4): If you 'speak the truth' you assert something true that you also take to be true. (The injunction '*Sag die Wahrheit*!' is a request for sincerity.)

I shall now sketch two formal reconstructions of the Self-Confessed Liar under which it is logically valid and leads to an inconsistent conclusion. The key idea of the 'necessitating' reconstruction is that the proponent of the argument takes for granted that the assumptions in its second and in its fourth line are not only contingently true. (P:/It is possible that; N:/It is necessary that; L*x*/*x* is a liar; A*x*/*x* admits to being a liar; T*x*/*x* speaks the truth).

1	(1)	$P:\exists x\,(Lx \,\&\, Ax)$	assumption
2*	(2*)	$N:\forall x\,((Lx \,\&\, Ax) \rightarrow Tx)$	assumption
1, 2*	(3)	$P:\exists x\,(Lx \,\&\, Tx)$	(1), (2*); (modal) logic
4*	(4*)	$N:\forall x\,(Tx \rightarrow \neg\, Lx)$	assumption
1, 2*, 4*	(5)	$P:\exists x\;(Lx \,\&\, \neg\, Lx)$	(3), (4*); (modal) logic

The key idea of the 'de-modalizing' reconstruction is that utterances that sound modal sometimes are not to be understood as modal. Take the sentences 'Triangles always have three angles, and they can be equiangular' and 'Not all swans are white—swans can also be black'. In both cases, the pseudo-modal part is to be understood along the lines of *Some S are P*.[12]

1*	(1*)	$\exists x\,(Lx \,\&\; Ax)$	assumption
2	(2)	$\forall x\,((Lx \,\&\, Ax) \rightarrow Tx)$	assumption
1*, 2	(3*)	$\exists x\,(Lx \,\&\, Tx)$	(1*), (2); logic
4	(4)	$\forall x\,(Tx \rightarrow \neg\, Lx)$	assumption
1*, 2, 4	(5*)	$\exists x\,(Lx \,\&\, \neg\, Lx)$	(3*), (4); logic.

[11] In 488: 9–15 Bolzano speaks of two syllogisms, but in 487/488 he formulates only (1), (2), (4), (5). For two syllogisms we need four premisses and two conclusions. He seems to take (3) as understood, which gives us two syllogisms in which the conclusion of the first is one of the premisses of the second.

[12] Cp. Russell (1918, p. 231): 'One may call a propositional function possible, when it is sometimes true.' By '"Fx" is sometimes true' he means: $\exists x\,(Fx)$.

Bolzano quickly identifies the culprit who is responsible for the inconsistent conclusion in line (5/5*).[13] It is the assumption in (4*/4): it presupposes, as Bolzano puts it, a false conception of what it is to be a liar. According to this conception, a liar never says anything (that is true and that he takes to be true). If this were the case, then the assumption in (1/1*) would be false; which it is not. Just as a smoker (even a chain smoker) does not smoke all the time, a liar (even a notorious liar) does not always lie whenever he or she makes an assertion.

If we apply Aristotle's distinction between *simpliciter* and *secundum quid,* we can transform the fallacy into a valid argument with true premisses:

(1A)	Sometimes somebody who is simpliciter a liar confesses to being a liar.
(2A)	If somebody who is simpliciter a liar confesses to being a liar then he speaks the truth. Hence
(3A)	Sometimes somebody who is simpliciter a liar speaks the truth.
(4A)	If somebody speaks the truth then he is secundum quid not a liar. Hence
(5A)	Sometimes somebody who is simpliciter a liar is secundum quid not a liar.

Assumption (4) would be false if 'to speak the truth' were equivalent with 'to say something that is true', for a true statement can be a lie. This is a key point in Bolzano's explanation of the concept of a lie to which I shall turn in the next subsection.

Asserting something that is true does not suffice for speaking the truth (*die Wahrheit sagen*), one has to be sincere in one's assertion. Sometimes we use that phrase as if the truth of what is asserted were not even necessary for speaking the truth. In his autobiography Russell says about his younger friend G. E. Moore:

> He had a kind of exquisite purity. I have never but once succeeded in making him tell a lie, and that was by a subterfuge. 'Moore,' I said, 'do you *always* speak the truth?' 'No,' he replied. I believe this to be the only lie he had ever told. (Russell 1967, p. 64).

Here, the focus is *only* on sincerity. If Moore had given an affirmative answer, he would not have claimed to be infallible, i.e. immune against the risk of error. Russell wants to beguile his friend into giving an insincere answer by asking him whether he is always sincere. So he might as well have asked, 'Moore, do you *never* lie?' If Moore's answer really is his one and only lie then it is additional evidence for his purity, for his answer shows that he does not want to pride himself with his sincerity.

In a famous paper ('a sort of instant classic'; Mates 1981, p. 31) Saul Kripke gave an astounding reading to Russell's question:

> It is said [*hm!*] that Russell once asked Moore whether he always told the truth, and that he regarded Moore's negative reply as the sole falsehood Moore had ever produced … Russell … apparently failed to realize that if, as he thought, all Moore's other utterances were true, Moore's negative reply was not simply false but paradoxical. (Kripke 1975, p. 692, repr. 55. The little cough is mine, of course).

Is this not a bizarre misreading? Does Russell regard his friend's purity to consist in his never saying anything false (apart from his reply to Russell's cunning

[13] The argument would not be logically valid if it were depicted by lines (1), (2), (3), (4) and (5) in my reconstructions. The step from (1) and (2) to (3) is invalid (replace '*L*' by 'is a mountain', '*A*' by 'is golden' and '*T*' by 'is a valley'), and the step from (3) and (4) to (5) is also invalid (replace '*L*' by 'is golden' and '*T*' by 'is a mountain').

question)? Does Russell assume that Moore never makes a false assertion, not even when asked for somebody's phone number, address or the first name, not even when asked for an analysis of the concept of perception? Kripke did not entirely fail to notice the absurdity of his interpretation, for he adds this footnote to his remark: 'On an ordinary understanding (as opposed to the conventions of those who state Liar Paradoxes), the question lay in the sincerity, not the truth, of Moore's utterances.' I think we would do well to remain faithful to the ordinary understanding and to renounce the extraordinary understanding of those who state 'Liar' paradoxes, rather than apply it to Russell's entirely non-paradoxical description of a very remarkable personality. As for the conventions of those who state 'Liar' paradoxes, here is a more recent example: '[In this section] lying was equated with saying something false. Most logicians assume this definition of lying and I will be doing the same' (Eldridge-Smith 2004, p. 76).[14] I dare say that logicians make a glaring conceptual mistake if they assume that each and every false statement is a lie.

24.1.2 What Is a Lie?

Bolzano's answer to this question is not to be found in his *Wissenschaftslehre*. Before he began working seriously on his logic, he held the chair of the Philosophy of Religion at Prague University for 15 years until the Emperor personally saw to it that he was sacked. He was also the university chaplain who had to deliver a sermon, a so-called exhortation or *Erbauungsrede,* on each and every Sunday. In one of those sermons, he answered our question as follows:[15]

> You lie if, and only if, you knowingly and deliberately give occasion that another person accepts something you yourself take to be false because you have testified it to him/her. (Man lüget nur dann, dann aber auch immer, wenn man mit Wissen und Willen Gelegenheit gibt, daß unser Nebenmensch etwas, welches wir selbst für irrig halten, auf unser Zeugniß glaubt und annimmt.)

By '*x* testifies to *y* that *p* (*x bezeugt y, dass p*)' Bolzano means, as he explains at length in his lectures on the philosophy of religion (Bolzano 1834, pp. 80–84, original pagination), that *x* tries to bring it about that the following situation obtains: *y* believes that *p,* because *y* takes *x* to believe that *p*. We can codify Bolzano's account of lying as follows:

(Df. L)$_B$	*x* lies to *y* if, and only if, for some *p*,
(1)$_B$	*x* tries to bring it about that (*y* believes that *p,* because *y* takes *x* to believe that *p*) &
(2)	*x* believes that not-*p*.

[14] As a matter of fact, by 'assuming that lying involves saying something false' (75) he had not maintained an 'equation' but 'only' a false implication. (He continues: 'However, we may briefly consider another definition…' If that definition were correct every case of concealing something would be a case of lying.).

[15] In this subsection, quotations are from Bolzano (18.03.1810, pp. 294–295), unless otherwise indicated. For a more detailed analysis of the pertinent passages cp. my (1999, pp. 139–153).

As regards the second clause, Bolzano agrees with Augustine and Aquinas, with Kant and (as we shall soon see) with Frege: What matters for the question whether an utterance is a lie is not the actual truth-value of what is said, but only the speaker's disbelief:

> What is relevant for lying is only that we ourselves *regard* as false what we are saying to other people; it is irrelevant whether it actually *is* false. (*Nur darauf ... ob wir dasjenige, was wir vor Andern aussagen, selber für… falsch halten, kömmt es bei der Lüge … an; nicht aber darauf, was an sich selbst betrachtet … falsch ist.*)

Consider the case of Little Johnnie. He is afraid of the maths test this morning, so he wants to stay in bed. By groaning and moaning he makes his mother hasten to his room, and when she opens the door he whispers, 'I am ill, Mum.' She gets the thermometer, and to his utmost surprise it turns out that he is really ill. Nevertheless, Little Johnnie *lied* to his mother, did he not?[16] So, as regards its propositional content, a lie need not comply with a necessary condition for being a *deception*. In his sermon on the slogan *'Mundus vult decipi, ergo decipiatur* (The world wants to be deceived, so let it be deceived)*'*, Bolzano specifies two necessary conditions of deceiving (*täuschen*) (For references see Künne 1999):

If *x* deceives *y* then for some *p*,
(a) *x* brings it about that *y* acquires, or retains, the belief that *p* &
(b) not-*p*.

As the case of Little Johnnie shows, not all lies comply with (b), the *falsity* condition of deception. And brief reflection suffices to see that a lie does not have to satisfy (a), the *success* condition of deception either: After all, people are sometimes silly enough to produce lies that are just too gross for anyone to be taken in by them.[17]

According to Bolzano's account, in lying you *try* to deceive your addressee. (This is not to say that every liar tries to *cheat* the addressee*,* for, as Bolzano explains (for references see Künne 1999; or 2013, p. 27 f.), cheating (*betrügen*) is a special case of deceiving: *x* cheats *y* just in case *x* deceives *y* with the intent of causing harm to *y,* or of winning an unfair advantage over *y*. One can lie to somebody without having such sinister intentions.) Augustine also emphasized that the 'intent to deceive *(voluntas fallendi)'* is a constitutive feature of lying, and many recent

[16] More famous is Kant's example in his paper 'On a Supposed Right to Lie from Philanthropy' (1797): In order to save somebody's life you lie when asked for his whereabouts, but as it so happens what you say is true, and so his pursuers get hold of him. Sartre fleshed this out in his story 'Le Mur' (1939), set in the Spanish Civil War. Moore also insists, as against Russell and the Shorter Oxford English Dictionary, that what is said in a lie need not be false: (1948/1949, pp. 381–382). If you regard this to be a matter of course, have a look at the beginning of Beall and Glanzberg (2011): 'The first sentence in this essay is a lie. There is something odd about saying so, as has been known since ancient times. To see why, remember that all lies are untrue…'

[17] So Sissela Bok is plainly wrong when, in her widely read and generally rewarding book Lying: Moral Choice in Public and Private Life (1978), she classifies lying as a species of the genus deceiving.

authors agree with Augustine and Bolzano.[18] When you look closely at clause $(1)_B$ in Bolzano's definition, you see that here two deceptive intentions are declared to be necessary for lying. A compound sentence of the form '*A because B*' entails both '*A*' and '*B*'. So, firstly, the speaker *X* wants the addressee *Y* to have (to acquire or to retain) a false belief on the topic of the proposition that *p*. Let us call this the speaker's *thematic* deceptive intention. Secondly, *X* wants *Y* to have a false belief about *X*: He wants *Y* to regard him as sincere. Call this the speaker's *attitudinal* deceptive intention. Furthermore, according to $(\text{Df. L})_B$, the speaker wants to reach his thematic goal *via* reaching his attitudinal goal.

Several philosophers, of early and late, have denied that lying always involves a deceptive intent. Aquinas denied it (for reasons I find rather opaque; for reference and discussion see Künne 1999), and so did (for a a transparent reason I shall soon come to) some contemporary philosophers (Carson 2006; Sorensen 2007; Fallis 2010). None of them seems to be aware of the fact that Gottlob Frege, too, implicitly denied that the intent to deceive is a constitutive feature of lying. (Only Aquinas has an excuse.) In his most widely read paper, Frege defines lying, more or less *en passant,* as follows (Frege 1892, p. 37, note 8; cp. Frege 1969, p. 252; 1979, p. 234):

(Df. L)F	x lies (to y) iff for some p,
(1)F	x asserts (vis-à-vis y) that p &
(2)	x believes that not-p.

By adding the bracketed material I have tried to improve a bit on Frege's account. I think he is right in not requiring *all* assertions to have an addressee: When I murmur, 'Damned, I shall miss my plane', I give vent to my anger, and I make an assertion even if I am not talking to anyone. But an assertoric utterance cannot be a *lie* unless it has an addressee. Incidentally, some philosophers conceive of assertion in such a way that clause $(1)_F$ is only satisfied if clause $(1)_B$ is satisfied (Chisholm and Feehan 1977, p. 152). That is not a good idea, I think. Suppose the management of a hotel accuses a chambermaid of theft. At the end of the questioning she exclaims in desperation, 'I know that nobody will believe me, but I did not take that money.' Her utterance is coherent, and in its second half she *asserts* that she did not take the money.

Now, are those who do demand that the producer of a lie aim at deception *right*?[19] Let me begin with the *thematic* deceptive intention Bolzano imputes to every liar. The following example speaks against this imputation.[20] Suppose Bob, an elderly hippie, observes from his window two policemen who are inspecting his rather

[18] In his (p. 395, 490) Augustine does not yet commit himself to the claim that mendacium est enunciatio falsum enuntiare volentis ut fallat (a lie is a statement of a person who wants to assert a falsehood in order to deceive somebody), which he subsequently endorses in (p. 420, 537) and in (p. 422, 243). Moore (1948/1949, p. 381), Barwise and Etchemendy (1987, p. 3), Williams (2002, p. 96) and Sainsbury (2009, p. 127) agree.

[19] Figures of speech such as understatement and irony are not counterexamples to Frege's account of lying. If an utterance of the sentence '*p*' is an understatement, or ironical, then the speaker does not put forward as true what '*p*' literally expresses, hence he does not assert that *p*.

[20] I owe it to Lisa Grunenberg.

untended garden. To his great displeasure, he notices that they have just discovered behind one of the elder bushes the hemp (*Cannabis sativa*) he planted there last year. The bell rings, Bob opens the door, and one of the cops says, 'We are from the Drugs Section of the Police Department, and we are currently having a look at some of the gardens in this suburb. Do you by any chance have some hemp in the garden behind your house?' Bob answers, 'No', knowing very well that some cannabis grows in his garden. He hopes that the policemen will come to believe that he has no idea what kind of herbs grow in the rear of his garden. Bob *lied* to the policemen, did he not? But he had no thematic deceptive intention, for he knew only too well that they had already found out the truth about the herbs behind the bushes.

Of course, in Bob's case an *attitudinal* deceptive intention is still in place. But there seem to be lies that are barefaced: they involve no deceptive intent whatsoever, neither thematic nor attitudinal. Is this an empty, a necessarily empty concept? In some quarters, barefaced lies are called baldfaced, and the *American Heritage Dictionary of the English Language* defines baldfaced lies as undisguised lies. If lying necessarily involves intending to deceive, then lies have to be disguised, and an undisguised lie is a *contradictio in adiecto,* something like an unintentional murder or a married spinster. But is this notion really inconsistent (cp. Sorensen 2007)? A sworn-in witness in a murder trial might make a statement that she knows to be false because she fears that friends of the defendant would kill her if she were to reveal in the witness box what she knows about the crime. Before answering the decisive question of the prosecutor, this witness might be certain that nobody in the courtroom will believe what she is about to say *and* that everybody in the courtroom will be convinced that she herself does not believe it either, and yet she gives the answer she know to be false. So both thematic *and* attitudinal deceptive intent are absent. Nevertheless, in committing perjury she is lying. Hence, Bolzano's explanation of the concept of lying seems to be too narrow.

Even so, his account might be a good account of lying with deceptive intent:

(*)	x lies to y with deceptive intent iff for some p,
(1)B	x tries to bring it about that (y believes that p, because y takes x to believe that p) &
(2)	x believes that not-p.

But some emendation is still required. Consider the right-to-left half of this biconditional: If for some *p*, $(1)_B$ and (2), then *x* lies to *y* with deceptive intent. By endorsing this conditional, Bolzano takes a stance on a debate which was triggered by the fact that the personnel of the Holy Scriptures and of the Life of the Saints is not always beyond suspicion of lying to other people. Consider the following incident in Egypt AD 362. Emperor Julian the Apostate wants Athanasius, the vociferous Patriarch of Alexandria, to be put to silence once and for all. Athanasius narrowly escapes an attempt to arrest him. In disguise he leaves the town by boat. Soldiers are sent out to search for him, their galley passes his boat, and they call out to the peasant-like looking stranger, 'Where is the traitor Athanasius?' 'Not far from here', answers the man in the boat, and unchallenged he rows on while the galley

accelerates (Cp. Theodoret III.9, col. 1095). Obviously, what Athanasius *conveyed* to the soldiers is something he disbelieved, sc. that the wanted person was not at the place from where they were answered. But did he lie to them? What he asserted is something he took to be true. (After all, if somebody is here then *a fortiori* he is not far from here.) That is why Augustine, Aquinas and Frege, unlike Bolzano, would not call him a liar, and I think they are right.

The following anecdote about a person who is not much of a saint can serve to make the same point: 'Where are you going?', asks her husband, handing her the car key. 'I need a pair of summer shoes. Yesterday I saw some very nice ones in a shop in the city,' she replies, and off she goes to meet her lover. She believes what she asserts, let us assume, but what she *communicates* to her husband is something she disbelieves. Augustine and Aquinas would not exactly approve of her activities, but unlike Bolzano they would not say that she *lied* to her husband. As a matter of fact, I would not either.

Frege would say of the proposition Athanasius made the soldiers accept as true, and of the proposition the adventurous lady made her husband believe, that they are not the thoughts 'explicitly expressed' by their utterances but rather 'subsidiary thoughts (*Nebengedanken*)' insinuated by them. Frege rightly maintained that this distinction is 'important for the question whether an assertion is a lie, or an oath a perjury' (Frege 1892, p. 46 f.; cp. Künne 2010, pp. 447–451). Bolzano is clearly aware of the distinction between what is literally said and what is indirectly conveyed, but he declares it to be irrelevant for the question whether an utterance is a lie or not. He says in his exhortation (Bolzano 18.03.1810, p. 297 f.):

> If we were to choose our words artfully in such a way that—though they express a truth if literally understood—they are meant to be taken by our audience in such a way that they convey a falsehood, then we would comfort ourselves in vain with the thought that we are sincere, we would be liars. (*wofern wir unsre Worte, mit absichtlicher Kunst so auswählten, daß sie zwar ihrem buchstäblichen Sinne nach die Wahrheit aussagen, von unsren Zuhörern aber in einem ganz andern und falschen Sinne genommen werden sollten: dann würden wir uns vergebens unsrer Wahrhaftigkeit trösten, wir wären Lügner.*)

There is some truth in this: The Patriarch rowing away and the lady going astray are not *sincere* (*wahrhaftig*), but that does not entail that their utterances are lies. I think Bolzano is misled here by his justified moral aversion against the Jesuit casuistry that Pascal had so stridently criticized in his 'Lettres à un Provincial' (1656). Bolzano is right in thinking that deliberately conveying a falsehood by explicitly expressing a truth can be as wicked as explicitly expressing a falsehood that one knows to be a falsehood.[21] But of course, the possibility that an act of the former kind is as reprehensible as an act of the latter kind is not a good reason to classify acts of both kinds as *lies*.

I sense another problem with Bolzano's clause $(1)_B$. The indeterminacy of the phrase '*x* does *something* in order to…' allows him to maintain that 'each communicative action is such that it can make its agent a liar (*durch jede* [*sc. bezeugende*] *Handlung … kann man zum Lügner werden*)' (Bolzano 18.03.1810, p. 296). Is this

[21] Jennifer Saul (2012) emphasizes the same point.

correct? Ben's widow is alone in her flat, and she is afraid of the strange-looking fellow who is standing in front of her at the door of her flat. She tries to chase him away by shouting over her shoulder, 'Ben! Could you come down for a second?' Let us hope that her attempt at communicative deception is successful, but did she really *lie* to the stranger? I do not think so. An utterance has to be an *assertion* if it is to be a lie, and Ben's widow did not assert anything when she pretended to call him. Now if we alter Bolzano's first clause accordingly, we also get rid of the problem with merely insinuated messages, so I suggest the following emendation of a Bolzanian account of lying with deceptive intent:

	x lies to y with deceptive intent iff for some p,
(1)B+	x asserts that p in order to bring it about that (y believes that p because y takes x to believe that p) &
(2)	x believes that not-p.

A speaker can assert that *p* without pursuing the goal ascribed to *x* in $(1)_{B+}$: recall the utterances of the desperate chamber-maid and of the frustrated traveller. Viewed in the light of this definition, several people Bolzano would call liars are cleared of this charge. But of course, that does not imply that they are beyond reproach, for they are *not* cleared of the charge of insincerity (*Unwahrhaftigkeit*).

24.1.3 Four Expositions of an Antinomy: 45 BC, 200, 1150 and 1910

Bolzano thinks that the Self-Confessed Liar is the so-called *ψευδόμενος* of Eubulides, *for which Philetas of Kos so strenuously sought a solution that he died from exhaustion* (WL III 487: *…über dessen Auflösung sich Philetas von Kos zu Tode studirt haben soll*). It is almost certain that this identification is mistaken.

Aristotle neither uses the sobriquet, nor does he mention Eubulides. The earliest source in which they are brought together is fairly late. Diogenes Laërtius (*fl.* [?] AD 250) tells us of a contemporary of Aristotle, 'Eubulides of Miletus who also presented many dialectical arguments in interrogatory form'. He then gives a list of seven arguments, and *ὁ ψευδόμενος* is the first entry (Diogenes Laërtius, *Vitae Philosophorum,* II. 108).[22] As for the title of this argument, the supressed noun is '*ἄνθρωπος* (man)', as can be seen from the non-substantival nicknames of arguments that follow in that list, such as 'the Bald (sc. Man)'. *To that extent,* its customary translation as

[22] Cp. Döring fragment 64, pp. 19, 106–110. According to Plato's Phaidon (59 C) Euclides, a native of Megara on the Isthmus of Corinth, was with Socrates when he died. 'His followers', so we are told in Vitae II.106, 'were called Megarians after him, then Eristics, and at a later date Dialecticians … because they put their arguments into the form of question and answer.' Eubulides is said to be one of his followers.

'the Liar' is correct. (I shall soon explain my reservation.) Diogenes Laërtius does not say a word about the content of any of those seven arguments.[23]

The anecdote about the poet and grammarian Philetas (or Philitas), a native of the Aegean island of Kos (*fl.* 300 B.C.), advises caution when dealing with the argument. The epitaph on his gravestone read, 'Stranger, I am Philetas. Of the arguments the Pseudómenos / Ruined me and the nocturnal brooding over<what it says>.'[24] In 1908, an Oxford don explained and translated this epitaph, tongue in cheek, as follows (Stock 1908, p. 36):

> [H]e grew thin and died of the Liar, and his epitaph served
> as a solemn reminder to poets not to meddle with logic–
> *Philetas of Cos am I,*
> *'Twas the Liar who made me die,*
> *And the bad nights caused thereby*.
> Perhaps we owe him an apology for the translation.

(Perhaps you do, Dr. Stock. In your translation the epitaph sounds as if it were a fragment of a limerick.)

This very anecdote makes Bolzano's identification of the Self-Confessed Liar with the Pseudómenos look very implausible. Can the attempt to spot the mistake in the fallacious argument (1)–(5) be responsible for Philetas' insomnia and premature death? The sentence 'I am a liar' can be used by a liar when he admits that he is a liar. It expresses as many truths as there are liars, so it expresses quite a lot of truths. It is *logically* absolutely harmless, and Bolzano recognized its logical harmlessness.

What he did not recognize is the logically explosive force of the antinomy that nowadays circulates under the label 'The Liar'. Let us examine the earliest formulation that has come down to us. Cicero makes the title character in his dialogue Lucullus (45–43 BC) criticize the scepticism of the New Academy. In the course of

---

[23] More about Diogenes Laërtius on Eubulides and about the ancient name of the argument in Appendix I below.

[24] Thus Athenaeus, a native of the Egyptian town Naukratis, (fl. AD 190) in his Δειπνοσοφισταί (Scholars at a Banquet, bk. IX, ch. 64, 401e). For the Greek text of the epitaph see Rüstow (101), Döring (110), Hülser (vol. 2, pp. 834–835).<Between the angle brackets the Greek text is opaque.>A much later testimony for the anecdote is contained in the article on 'Philitas' in a Byzantine encyclopaedia that has come down to us under the names Suda and Suidas (c. 970): The passage is quoted in Rüstow (101). At least since the middle of the seventeenth century, this anecdote enjoyed ever-growing popularity. For Pierre Gassendi, the Pseudómenos and the other 'Megarian' arguments listed in Vitae were 'nothing but captious reasoning (nihil aliud quam captiosae quaedam rationes)', and he did not miss out on the epitaph. Pierre Bayle remarked: 'It is not an exaggeration but a literal truth that Eubulides' inventions were deadly sophisms (des Sophismes à tuer les gens)' (Gassendi 1658, p. 40; Bayle 1697, p. 415). Presumably Bolzano knew the anecdote through Hegel and/or Fries. In Hegel's Lectures on the History of Philosophy (1833, p. 136), the epitaph is quoted from Isaac Casaubon's Greek–Latin edition of Athenaeus' book (Heidelberg 1597, Lyon ^{4}1664), and Hegel also mentions Suidas as a source of the anecdote. Bolzano refers to this volume of Hegel's works in WL I 39. Hegel's opponent Jakob Friedrich Fries also relies on Athenaeus and Suidas when he reports in his *System der Logik* that Philetas 'sich am ψευδόμενος des Eubulides…zu Tode studirt haben soll' (Fries ^{1}1811, 470=^{3}1837, p. 356). As we saw above, Bolzano used the very same words, and he knew Fries' book: In various parts of the WL it receives critical attention, a copy of the second edition belonged to his private library.

his attempt to defuse this criticism Cicero wants to topple the Stoic definition of a declarative sentence (in this text he translates ἀξίωμα as ecfatum):

> Now what about the following declarative sentences, are they true or false?
> [1?] If you say that you are lying and if you thereby say something true, you are lying
> (*si te mentiri dicis idque verum dicis, mentiris*)
> [2?] <and: If you say that you are lying and if you *are* lying, > you say something true
> (*<et: si te mentiri dicis idque mentiris, >verum dicis*).
> Of course, you will say that they are inexplicable (*haec…inexplicabilia esse*)… But now I ask: if they are not explicable and if no criterion (*iudicium*) can be found by means of which the question whether they are true or false could be answered, what becomes of your definition of a declarative sentence as something that is true or false? (*Lucullus* (=: *Academica priora,* Buch II), § 95).[25]

(I shall soon explain the question marks that are related to the word shaded in grey in my provisional translation.) Obviously, Bolzano cannot have had this text in mind when he constructed the Self-Confessed Liar.[26] Cicero does not mention Eubulides,[27] but apparently he takes the Greek nickname to stand for the pair of *inexplicabilia* in the above text, and he also coins the Latin counterpart to this label: 'How the *Mentiens* that they call the *Pseudómenos* (*mentientem quem ψευδόμενον vocant*) is to be solved…the dialecticians will tell us' (Cicero, *De divinatione* II.11, in Cicero 1991, p. 142). A century later, Seneca dismisses the Pseudómenos as a nuisance when he writes to his friend Lucilius (AD 64): 'Why do you hold me up with the *Pseudómenos* as you yourself call him (*quem tu ipse pseudomenon appellas*),

[25] The Latin text is gappy in all codices. Like Hülser fr. 1212 and Cavini (1993, p. 91) I endorse the supplementation proposed by Otto Plasberg in 1908 and repeated in his edition [Bibl.] Cicero-(B) 73. Rüstow (89) inserts between the angular brackets only the word 'an (or)', as did James Reid in Cicero-(A); Olof Gigon also accepts this conjecture: Cicero-(C) 210 f. Spade (1973, II. p. 298 f.) uses the text in Cicero-(A). (He mistranslates the conjunctive antecedent of [1] as 'If you lie and speak that truth,' and this mistake is repeated in Spade/Read § 1.1, where this ill-formed English sentence is praised as a 'fairly clear formulation'.) As Walter Cavini (loc. cit.) points out, Cicero's use of the plural 'inexplicabilia' is incomprehensible in the (A/C) text: Only the conditionals in Plasberg's reconstruction give us two 'inexplicable' sentences. A first-person variant of the reconstructed conditional [2] has been found in a dictionary by Placidus [fl. (?) AD 550]: 'si dico<me>mentiri et mentior, verum dico' (Rüstow p. 102; Hülser fr. 1217). The way Cicero continues after the passage quoted above (Acad. Pr. II. pp. 96–98) is logically very unsatisfactory (cp. Rüstow pp. 88–91). At three other places, Cicero just mentions the Pseudómenos (always in tandem with the Sorites: Acad. Pr. II.147, De div. II.11) (see below) and Hortensius [in Cicero-(C) 64] (vgl. Rüstow pp. 88–91). Both Hegel (op. cit. 135) and Fries (loc. cit.) take 'the Pseudómenos of Eubulides' to be the antinomy in Cicero's version. In this respect, they seem to come closer to the historical truth than Bolzano.

[26] Bolzano draws on Cicero's book in the section on fallacies when he discusses the *Sorites* (pp. 494–495). (The title of the book is somewhat unstable. He calls it *Acad*[emicae] *quaest*[iones], just as Bayle (1697) *passim,* Hegel (loc. cit.) and Fries (loc. cit.) did. In vol. IV.1 of the Cicero edition by Johann Caspar von Orelli (Zürich 1828) that stood on his bookshelves, the work has the title *Academica priora*. Its first book is often called *Lucullus*.)

[27] Pace Eldridge-Smith (2004, p. 78): 'We know about Eubulides' formulation indirectly from Cicero.'

about whom so many books have been written?' (Seneca, *Epistula* XLV.10).[28] The many papyrus scrolls he alludes to are all extant: according to Diogenes Laërtius, three of these works were written by Theophrastus, Aristotle's successor in the Peripatetic School, and no less than 21 by the Stoic Chrysippus (*Vitae*, V.49 and VII.196–197).[29]

If we transform the indirect speech in Cicero's conditionals into direct speech, the following sentence becomes visible as the source of the Pseudómenos:

(lat.) (ego) mentior.

Retranslated into the language of the original presentation that is not extant:[30]

(gr.) (*ἐγὼ*) *ψεύδομαι*

Apparently, we have here reached a point of convergence with Bertrand Russell's exposition of the antinomy in the Introduction to *Principia Mathematica* (1910): *The simplest form of the* [*antinomy*] *is afforded by the man who says* '*I am lying*'.[31] And now it seems clear how the Greek sobriquet of the antinomy is to be translated: 'the Man Who Is Lying'. But sometimes appearances are deceptive, and this is a case in point. The verbs in (lat.) und (gr.) are equivocal, and in the same way (Cp. Barnes 2007, pp. 7–8). For each of these sentences, there are two correct translations into Russell's language:

(*L*) I am lying
(*F*) I am saying something false.

Note first that neither of these sentences expresses what is expressed by 'I am a liar'. In uttering (*L*) or (*F*) you do not say what kind of a person you are but rather what you are currently doing. Secondly, (*L*) and (*F*) are not equivalent, and there are two reasons for this nonequivalence: Obviously, you can say something that is false without lying, and (as we saw in Sect. 24.1.2) you can lie and inadvertently say something that is true. Thirdly, as reading (*F*) shows, in Greek *and* Latin there is a

[28] Gassendi (1658) 40 quotes this remark, and Bayle applaudes it: 'It is good to see how Seneca makes fun of those who waste their time with such useless quibbles (vaines subtilités)' (Bayle 1697, p. 415). (Spade/Read draw from Seneca's remark the false conclusion that the antinomy had '[the] Greek name "pseudomenon"' (§ 1.2). When Cicero and Seneca use the form '–on' in their remarks, they only follow a grammatical rule that requires the participle to be put into the accusative in those contexts. The gender of the nickname is masculine. For a satire on this see Appendix I.)

[29] Cp. Rüstow (54, pp. 60–86) and Cavini (1993, p. 102) on whose count I rely in the case of Chrysippus.

[30] But compare the two commentators on Aristotle I shall quote below.

[31] Already in Russell (1908, p. 59), repeated verbatim in Principia (1910, p. 63). The earliest occurrence might be that in Russell (1906, p. 197). Quine echoes Russell's rendering of 'the ancient paradox' in his (Quine 1962, p. 7).

semantically *atomic* expression that translates 'to say something that is false'. (Some scholars try to get a bit closer to this by rendering (gr.) and (lat.) as 'I am speaking falsely' (e.g. Cavini 1993). But would one not use this sentence rather for criticizing one's own speech as ungrammatical?) In *Greek,* there is also a semantically atomic expression that translates 'to say something that is true': The laconic counterpart to 'I am saying something that is true' is '*ἀληθεύω* [alētheúō]'. (This is sometimes rendered as 'I am speaking truly' (e.g. Cavini 1993.) But does not that force the adverb into a service for which it is never employed in nonphilosophical English?).

In the present context, reading (*F*) of '*mentior*'/'*ψεύδομαι*' is to be preferred. In any case, the special intentional profile of the person who is lying, let alone that of the person who is lying with deceptive intent, is entirely irrelevant to the antinomy.[32] A wooden translation is a lesser evil, I take it, than a graceful mistranslation: the Greek nickname of the antinomy should be rendered by 'The Man Who Is Saying Something False' or 'The Falsehood-Teller' rather than by 'The Liar'. The latter title far better suits the fallacy that *Bolzano* called by that name. This is just for the record: The mislabelling has become routine by now, so it is most unlikely that my admonition will ever be heeded in the literature on the antinomy. Be that as it may, henceforth I shall refer to the so-called Liar Paradox as the *antinomy of falsity*, briefly: F-antinomy,[33] and I shall call sentences that are the source of the F-antinomy *F-sentences*. I take a *paradox* to be a statement that seems to be obviously false, hence 'beyond belief (*παρὰ δόξαν*)',[34] although it can be derived—from premisses whose truth is apparently as obvious as the falsity of the conclusion—by a reasoning whose validity seems to be equally obvious (Sainsbury 2009, p. 1).[35] This dovetails with Aristotle's characterization of the deadlock (*ἀπορία*) caused by arguments that seem to prove *παρά δοξα*: 'Thought is enchained (*δέδεται*), being unwilling to rest because it cannot accept the conclusion reached, yet unable to go forward because it cannot refute the argument' (Aristotle, *EN* VII. 2: 1146^{a}24–27). I clas-

[32] According to the interpretation I offered in Sect. 24.1.1, it is also irrelevant for the point Aristotle wants to make in SE 25 at [d], sincerity is the issue only in [f].

[33] 'Semantical antinomy' has a more familiar ring, and it also improves on 'The Liar'. But strictly speaking, '*x* is false' is not a semantical predicate unless it is applied to sentences, and the title of the antinomy should not commit us to this. Following Łukasiewicz (1915), Alfred Tarski presents the F-antinomy as a semantical antinomy. He calls it 'the antinomy of the liar', but he never muddies the water by using terms such as 'lie' or 'liar' (Tarski 1932, § 1; 1944, § 7; and 1969, part I).

[34] For '*πὰϱ* δύναμιν, beyond one's strength' LSJ refers to the Iliad, for '*παϱὰ* δόξαν, contrary to belief' to Thucydides. Cp. also Aristotle, Top. I.11: 104b19, 24, 34. Cicero defines paradoxes as sentences 'that are amazing and that go against everyone's beliefs (mirabilia contraque opinionem omnium)', and he reckons with the possibility that many of them are true (*Paradoxa Stoicorum*, Pref. § 4). For the author of *Paradoxien des Unendlichen*, paradoxes are 'propositions that sound strange (befremdlich klingende Sätze)' and that may have the 'appearance of being self-contradictory (Schein des Widerspruches)' although many of them are true (Bolzano 1851, §§ 1, 28). Cp. Quine (1962) on 'veridical, or truth-telling, paradoxes' like that of the village barber who allegedly shaves all and only those villagers who do not shave themselves.

[35] Quine (1962) only begins with such a characterization (note 37).

sify a paradox as an *antinomy* if it is self-contradictory.[36] In a transferred sense, as *totum pro parte* ('America' for the 'USA'), I apply both terms also to arguments that sustain paradoxes (antinomies) in the literal sense (Eldridge-Smith 2004, p. 83).[37]

Cicero's conditionals 'si te mentiri dicis idque verum dicis, mentiris' und 'si te mentiri dicis idque mentiris, verum dicis' deserve a better translation:

[1] If you say that what you are saying is false and if you thereby say something true then what you are saying is false

[2] If you say that what you are saying is false and if you *are* saying something false then what you are saying is true.

These conditionals are substitution instances of two intuitively immensely plausible (schematic) principles:

[P1] If (*x* says that *p*, and *x* thereby says something true), then *p*.

[P2] If (*x* says that *p*, and *p*), then *x* says something true.

Harmless instances are 'If Plato says that Socrates is wise and if Plato thereby says something true, then Socrates *is* wise' and 'If Plato says that Socrates is wise and if Socrates *is* wise, then Plato says something true'. The conjunction of [P1] and [P2] is as plausible as the more laconic denominalization schema:[38]

(Den) It is true that *p*, if, and only if, *p*.

Cicero's conditionals [1] and [2] seem to show that the application of the apparently unassailable schematic principles [P1] and [P2] to a speaker *X* who says 'I am saying something that is false' has the paradoxical consequence that what *X* says is true if, and only if, what *X* says is false (Spelt out in Cavini 1993, pp. 98–102).

Did *Aristotle* ever confront the F-antinomy in his writings? If he did, then it was in the passage [**b**]-[**f**] in Chap. 25 of his *Sophistical Refutations* that we scrutinized in Sect. 24.1.1. Apparently, in the Middle Ages most philosophers who discussed the antinomy were convinced that this passage was concerned with it (Spade/Read

[36] Thus, Church (see below, Sect. 24.1.4) and Quine (1962) p. 5.

[37] In Quine (1962), *from p. 2 onwards,* and in John Mackie (1973) 238, 270 the use of 'paradox' that I take to be secondary is regarded as primary. (A third characterization should at least be mentioned: According to Nicholas Rescher (2001, p. 6 f), Roy Sorensen (2003, p. 6) and esp. William Lycan (2010) a paradox is an inconsistent set of statements each of which is extremely plausible.)

[38] Künne (2003, 18 et passim, see 488). On p. 151, I discuss its Aristotelian precursor in Int. 9 (18a39–42). [The most thorough and exhaustive study on Aristotle's conception of truth is Crivelli (2004).] Cavini (1993, p. 92) calls [P1] 'the [first half of] the Stoic "paratactic" Equivalence Thesis'. He gives good reasons for using the sobriquet 'Stoic'. But by 'paratactic' he cannot mean paratactic, since the schema is obviously hypotactic, and the term 'Equivalence Thesis' does not distinguish (Den) from the disquotation schema '"*p*" is true, iff *p*' (Künne 2003, p. 14 et passim, see 488).

§ 1.3 u. § 2.1),[39] and some more recent authors share this conviction. We are told that the antinomy is 'discussed' in [**b**], that it is 'explained' in [**b**]–[**e**], or that it '*kommt ausdrücklich zur Sprache*' in [**d**]–[**e**].[40] Let us have another look at the latter part of the passage, for it is here that Paolo Crivelli sought support for an even bolder hypothesis:

> [**d**] The argument is similar when we turn to the question whether it is possible for the same man at the same time to say something that is false and to say something that is true. [**e**] But since it is not easy to see whether we have a case of *simpliciter* saying something true or a case of *simpliciter* saying something false, here the issue seems to be difficult (*δύσκολον φαίνεται*).

According to Crivelli, (a) Aristotle discusses in these lines the antinomy that arises from utterances of 'What I am saying is false' *and* (b) tries (in vain) to solve it by letting such utterances fall into the truth-value gap. In the end, Crivelli himself concedes: 'None of the above considerations is decisive' (Crivelli 2004, pp. 31–34, 139–151, here 147). I think that only one of his considerations is above the suspicion of being merely speculative. What does the special 'difficulty' that according to [**e**] distinguishes the question raised in [**d**] from those mentioned before consists in? This could indeed be easily explained if the pertinent utterance were the utterance of an *F*-sentence (Op. cit. 145, building on Rüstow 50–51). But can one explain it *only* if one makes this assumption? The question whether somebody who asserts, e.g. 'Achilles is a *lion,*' or 'The *Greeks* won the Battle of Marathon' said something that is true without qualification or false without qualification is also more difficult to answer than the question (Cp. *SE* 5: 167^{a}10–12) whether an Ethiopian beauty with pearly white teeth is *simpliciter* black or *simpliciter* white. I do not think that Aristotle, for one reason or another, did not succeed in *SE* 25 to solve the F-antinomy (as do Rüstow pp. 49–55; Bocheński, pp. 152–153; Sorensen 2003, p. 198; Crivelli 2004, p. 151), for I cannot see that the text contains a strong indication that he is adverting to the antinomy at all.

Two centuries or more after Cicero, we find an exposition of the F-antinomy and a proposal for its solution in the commentary on Aristotle's *Topics* that was written by Alexander of Aphrodisias (*fl.* AD 200), the most important Greek commentator on Aristotle's works. The context is this. In *Topics* II.7 Aristotle sketches a refutation strategy of the following kind: If an object *a* is said to be *F*, inquire whether there are any two properties *x* and *y* such that: *a* would have to have both *x* and *y* if *a* were *F,* but *x* and *y* stand in contrary opposition to each other. If there is such a pair, the claim that *a* is *F* is refuted (*Top*. II.7, 113^{a}24–25). Now Alexander explains this strategy by exemplification:

> [Aristotle] himself uses as example a thesis about [Platonic] forms [*Top.* II.7: 113^{a}26–32]… But one can use this *topos* also in order to refute (*ἀναιρεῖν*) the thesis that
> *I am saying something false* (*ἐγὼ ψεύδομαι*)

[39] Savonarola (Sect. 24.2) does not regard the F-antinomy as a case of the fallacia secundum quid et simpliciter, which he discusses two sections after his examination of the antinomy (CL 154–155).

[40] The quotations are, in this order, from Rescher (2001, p. 199), Sorensen (2003, p. 197) and Döring (216–217).

> is a declarative sentence (*πρότασις*); for if it is a declarative sentence then it is, as we shall see, both true and false, But that is impossible, for these properties stand in contrary opposition. Hence 'I am saying something false' is not a declarative sentence. For if one were to concede that it *is* then the principle would be refuted that every declarative sentence is either true or false. For if the latter principle is correct this declarative sentence is also either true or false. But whatever we assume, it entails the contrary. If we assume that 'I am saying something false' is true, then it seems to follow that the speaker is saying something false, for then by saying about himself that he is saying something false he would say something true. On the other hand, if we assume that 'I am saying something false' is false, it follows that the speaker is saying something true, for then by saying about himself that he is saying something false he would say something false, hence he would say something true. (Alexander of Aphrodisias 1891, p. 188, 19–28; *Hülser* fr. 1183)[41]

Strictly speaking, Alexander does not establish that his *F*-sentence is both true *and* false, if it is a declarative sentence, but only that it is true just in case it is false. (In Sect. 24.2.1 we shall see how the conjunction can be derived from the biconditional and the principle of bivalence.) Unlike Cicero, Alexander sees the principle that every declarative sentence is either true or false not (or not primarily) threatened by the conditionals [1] and [2] but rather by the *F*-sentence that these conditionals are about. In the next section we shall reencounter his proposal to dissolve the paradox by denying the *F*-sentence the status of a declarative sentence.

There is not the slightest hint in the above text that Alexander takes the argument to be due to *Aristotle*. On the contrary, he presents the F-antinomy as *his* example for an argumentative strategy to which a *topos* can be applied. Aristotle's conception of fallacies due to the neglect of a necessary qualification plays no role whatsoever in Alexander's argument.

Thousand years later, the F-antinomy also entered commentaries on the *Sophistical Refutations*. In a Byzantine commentary written by Michael of Ephesus [*fl.* (?) 1150], a paraphrase of *SE* 25 is followed by a very brief dialogical postscript on the question Aristotle had posed in [**d**]:

> Can one at one and the same time say something that is false and true?—No.—But look, if somebody says, 'I am saying something false', he does say at the same time something that is false and true (*ὁ λέγων 'ἐγὼ ψεύδομαι' ἅμα καὶ ψεύδεται καὶ ἀληθεύει*). So it is false that nobody can at the same time say something that is true and false. (Michael Ephesios 1898, pp. 171, 17–20. Cp. *Döring* 110)[42]

[41] (In Hülser's translation, one occurrence of the verb '' is rendered by 'etwas Falsches sagen' and all others, unfortunately, by 'lügen'.) Alexander's account of the F-antinomy seems to be the one and only ancient contribution to the debate that escaped Rüstow's attention. That is why the text is hardly ever mentioned in the literature on the antinomy. In 1988, Karlheinz Hülser made it easily available in his collection of fragments on the dialectics of the Stoics. But it is only in Cavini (1993, p. 89, 107) that I found a reference to this important document.

[42] I take Rüstow's word for it that Karl Praechter has finally demonstrated that not Alexander Aphr. but Michael Eph. is the author of this commentary (Rüstow 1911, p. 630; referring to Göttingische Gelehrte Anzeigen 1906, pp. 861–907). In ignorance of the passage in Alexander that is thousand years older, Rüstow says about Michael's PS that here it happened 'das erste…Mal, daß der Lügner in seiner absolut strengen und scharfen Form erscheint'. Scholz (1937, p. 264) and Mates (1981, p. 16) echo this venial misjudgement. In Mates (1981), the key statement in Michael's PS is first rendered, without reference to him, as 'When a man says "What I am now saying is false,"

The equanimity with which the affirmative answer to the question raised at the beginning is accepted at the end makes one wonder whether Graham Priest had a precursor in Constantinople.[43] The Byzantine commentator does not make (here) any attempt to take the sting out of the bold Yes by distinguishing between *secundum quid* and *simpliciter*.[44]

Let us return to Russell's exposition of the antinomy (and not for the last time). So far I did not let him finish: *[Its] simplest form…is afforded by the man who says 'I am lying'; if he is lying, he is speaking the truth, and vice versa* (loc. cit.). If the speaker is lying then he performs the sort of action that he ascribes to himself when he says:

(*L*) I am lying,

and if you *are* doing what you claim to be doing, then what you say is true. Does a person who claims to be doing *X* and who actually does *X eo ipso* 'speak the truth'? Strictly speaking, no, for the latter requires also that he *believes* that he is doing *X*. So let us replace 'speaks the truth' by the attitudinally neutral predicate 'says something true'. This gives us one half of Russell's biconditional: If the utterer of (*L*) is lying then he is saying something that is true. What about the other half of Russell's biconditional? If the speaker who uses the self-ascription (*L*) *says something true,* then he performs the very action that he ascribes to himself in uttering (*L*). This action consists in lying. Hence, if the (*L*)-utterer is saying something that is true, then he is lying. So far, so good. But not good enough, for (as G.E. Moore pointed out) the biconditional we have now obtained is not paradoxical. The impression that it is paradoxical is due to the false presupposition that you lie only if you say something that is false. For then the biconditional entails that by uttering (*L*) you say something true just in case you say something false by uttering (*L*) (Moore 1948/1949, p. 382).[45]

---

what he says is both true and false' (6) and then, with reference, as 'The man who says "I am lying" is both lying and telling the truth' (16). By now the reader knows where my preference lies.

[43] On Priest's position see Sainsbury (2009, p. 150–158) and Priest and Berto (2010). (Priest gave this position the odd name 'dialetheism', and he does not suffer from linguistic scruples when he calls contradictions 'dialetheias'.)

[44] This is done (in the most implausible way) by an even later Byzantine commentator (Hülser fr. 1218; cp. Rüstow, p. 107). Several authors in the Latin West try the same way out for which they (falsely) claim Aristotle's authority (Dutilh and Read 2008; and above note 9 et note 40).

[45] Benson Mates makes the same blunder when he writes: 'If a man says "I am lying," what he says is true if and only if it is false' (Mates 1981, p. 18). If the man is right in what he says then he is lying, but that does not imply that what he says is false. Moore's critical remarks in his Notebooks in 1948/49 are a late manifestation of his long engagement with Russell's text: 'As for his Introduction to Principia Mathematica [and five other works by Russell], I have … lectured in detail on particular passages … on various occasions … at Cambridge. Of course … my lectures on what he has written have always been partly critical. But I should say that I certainly have been more influenced by him than by any other single philosopher' (Moore 1942, p. 16).

Moore noticed another defect in Russell's exposition of the antinomy (Moore 1948/1949, p. 383; cp. also his Moore 1962, p. 291. I have replaced his example by my own). Suppose John mutters 'I am lying', while he is writing in a letter to his girlfriend: 'I love you'. Mimicking Russell's formulation in *PM,* we can describe John's situation as follows: If he is lying (in his letter), he is saying something true (in his oral comment on his letter), and vice versa. No paradox is in sight. Two pages later in *PM,* Russell says about sentence (*L*): *When a man says 'I am lying', we may interpret his statement as: 'There is a proposition which I am affirming and which is false'* (Russell 1908, 61 = 1910, p. 65; for the first time, perhaps, in Russell 1906, p. 207). I do not think that we may do this, for this interpretation of (*L*) is a misinterpretation.[46] And if our letter writer were to utter the second sentence mentioned here, or simply:

(*F*) I am saying something false,

as a comment on his written declaration of love, again no paradox would arise. If we want to obtain a paradox we must put something like the following sentence into the speaker's mouth:

(F_{refl}) I am saying something false by uttering this (⇆) sentence.

Here and in the rest of this chapter the pair of arrows (that can sometimes be pronounced as 'very') asks the reader for a reflexive reading: The demonstrative description serves to refer to the *very* sentence in which it occurs. Surely, Russell and his ancient precursors *intended* the self-referential reading of (*F*). But Moore rightly insists on maximal explicitness in an area as tricky as this one.

Perhaps[47] Eubulides was the first philosopher to notice the logical morass we run into as soon as we try to answer the question: Does an utterance of (F_{refl}) express a truth, or does it express a falsehood? In view of the 'Liar' industry since the seventies of the last century, it is not hard to believe that Theophrastus and Chrysippus wrote as much about the F-antinomy as Diogenes Laërtius claims they did. Many failed nocturnal attempts to solve *this* problem may really have foreshortened poor Philetas's life. He might have lived longer if he had opted for that treatment of the antinomy which Russell called the 'March Hare's solution': 'Suppose we change the subject,' the March Hare [said], yawning. 'I'm getting tired of this' (Russell 1959, p. 77; Lewis Carroll, *Alice in Wonderland*, Chap. VII).

In his *opus magnum,* Bolzano, too, brooded over the F-antinomy. Unfortunately, he also got tired fairly soon. Fortunately, when he tries to cope with the antinomy, he does not muddy the water by talking of lies or liars. In His Subject Index, there

[46] Frank Ramsey even contends that one has to understand 'I am lying' in the sense of 'Things are not as I am right now saying they are' (Ramsey 1925, p. 48). The phrase 'may be interpreted as' is replaced in Russell (1918, p. 262) by 'means' and in Russell (1940, ch. 4) by 'i.e.'. On Russell and Ramsey cp. again Moore (1948/1949, pp. 382–383; 1962, p. 291).

[47] I give reasons for this reservation in Appendix I.

is *no* reference to that passage. As we saw, he misidentified the Pseudómenos as the fallacy of the Self-Confessed Liar.[48] But before I turn to the relevant passage in the first volume of *Wissenschaftslehre,* I want to dwell upon a famous dictum by an ancient Cretan and its twentieth century misreading.

24.1.4 A Cretan Denigrator of All Cretans: Epimenides in the Letter to Titus

It is high time that I complete my quotation from the Introduction to *Principia Mathematica*. So far I have kept back Russell's first statements on the F-antinomy:

> The oldest contradiction of the kind in question [i.e. of the kind to be shielded off by the 'Theory of Types'[49]] is the Epimenides. Epimenides the Cretan said that all Cretans were liars, and all other statements made by Cretans were certainly lies. Was this a lie? The simplest form of this contradiction is afforded by the man who says 'I am lying'; if he is lying, he is speaking the truth, and vice versa. (Russell 1908, 59 = 1910, p. 63)[50]

Russell alludes here to a passage in a letter that Paul the Apostle is alleged to have written to his collaborator Titus whom he had 'left in Crete … to appoint presbyters in every city' (*Tit.* 1: 5). Since the early nineteenth century, scholars have argued that the 'Letter to Titus' is pseudepigraphical, that it was written around Anno Domini 100, several decades after Paul's death (Cp. Schnelle 2007, pp. 369–374, 388).[51] The author—I shall call him Ps[eudo]-Paul—warns the recipient not to trust the inhabitants of the island, and he quotes the testimony of a very ancient witness:[52]

---

[48] By contrast, when Friedrich Kambartel included the entry 'Lügner (Antinomie)' in his useful subject index (cp. Bolzano 1963, p. 376), he referred the reader, reasonably enough, to the entirely 'liar-free' passage to be examined in Sects. 24.2 and 24.3 of this chapter.

[49] Different versions of this theory are explained and discussed in Copi (1971).

[50] Russell brought Epimenides already in a letter to Philip Jourdain (28. 04. 1905) and in (Russel 1906, p. 196) into the play: 'I come now to…the case of insolubilia. Such paradoxes have been known ever since the time of Epimenides the Cretan, assuming him to have really said that all Cretans are liars.' He borrows the medieval term 'insolubilia' (that we met in Sect. 24.1.2 and that we shall encounter again in Sect. 24.2.1) from a dictionary article signed 'C.S.P.', from which he also quotes with consent. But the author of that article, Charles Sanders Peirce, neither refers to Epimenides, nor does he speak of liars or lies. He explains the term 'Insolubilia' by means of an example: 'Given the following proposition: This assertion is not true; is that assertion, which proclaims its own falsity, and nothing else, true or false?' (Peirce 1901).

[51] For a brief summary of the arguments, see Bornkamm (1969) 245. Two authentic Pauline letters, Ad Corinthios II and Ad Galatas, testify that a man named Titus really was an associate and confident of the apostle.

[52] Both in the Vulgata and in the Nova Vulgata the text runs as follows: *Dixit quidam ex illis proprius ipsorum propheta: 'Cretenses semper mendaces, malae bestiae, ventres pigri.' Testimonium hoc verum est*. For two reasons, I do not follow the translation in The King James Bible [1611], Oxford Standard Text (1769) which is: 'One of themselves, even a prophet of their own, said, The Cretians are alway liars, evil beasts, slow bellies. This witness is true.' Firstly, the word 'Cretians' is no longer in use (and it looks too similar to 'Cretins', anyway). Secondly, at least in modern

[*Tit*. 1: 12] One of [the Cretans], a prophet of their own, said,
Cretans are always liars (*Κρῆτες ἀεὶ ψεῦσται*),
evil beasts, lazy gluttons (*κακὰ θηρία, γαστέρες ἀργαί*).
[13a] This testimony (μαρτυρία) is true.

'For this reason,' Ps.-Paul continues, 'you must rebuke them sharply' (*Tit.* 1: 13b). (One may reasonably wonder whether he was 'insensitive to the insult he was inflicting on the Cretan brethren by the use of so devastating a quotation' (Gealy 1955, p. 531).) Who was the Cretan who cast his strong reprimand of all Cretans in the mould of an hexameter?[53] In the epistle he remains anonymous, but hundred years later Clement of Alexandria, in his *Miscellanies,* identifies Ps.-Paul's witness as the presocratic poet and soothsayer *Epimenides,* apparently a native of Knossos (*fl.* [??] 500 BC).[54]

Unfortunately, Bolzano never commented on the Epimenides passage.[55] Ever since Russell alluded to it, reference to Epimenides has become endemic in the literature on the F-antinomy. In 1947, Alexandre Koyré published a monograph on this antinomy under the (doubly misleading) title '*Épiménide le menteur*'.[56] Quine discusses the F-antinomy under the heading 'The Paradox of Epimenides' (Quine 1962, pp. 6–10; cp. Quine 1953, p. 133; Eldridge-Smith, 'The Cretan Liar Paradox' (2004)). Kripke says about this antinomy: 'as is well known, [it] arises in a New Testament context' (Kripke 1975, p. 690/repr. 53). (The verb in 'well known' is factive.) Rescher says about the F-antinomy that 'its notoriety was such that even St. Paul adverted to it' (Rescher 2001, p. 200). And Alan Ross Anderson recommends us '[to] bow to St. Paul and to a long tradition according to which "the Liar Paradox" and "the Epimenides paradox" meant the same thing' (Anderson 1970, p. 3). How 'long' is this tradition? Anderson cannot possibly mean that it begins in the *New Testament*. In medieval treatises on *Insolubilia,* scholars assure us, the

English, 'witness' has a reading that is unwelcome here: person who can give testimony (the other differences are irrelevant for the topic of this chapter).

[53] Just in case you need help: | ***Krē****-tes a-* | ***ei*** *pseus-* | ***tai*** *ka-ka* | ***thē****-ri-a* | ***gas****-te-res* | ***ar****-gai* |.

[54] Clement [fl. 190], Στρώματα oder Στρωματεῖς [lit. Patchwork Rugs] bk. I, ch. xiv, 59.2. In antiquity, many legends were told about Epimenides—the story about his waking up in a cave from a sleep that lasted several decades (Vitae I. p. 109) was taken up by Goethe in his play 'Des Epimenides Erwachen'. Jerome, in his commentary on the 'Letter to Titus' (written in 387), conveys the impression that he even knows the title of the (lost) work in which the hexameter quoted by Ps.-Paul occurred: 'Χρησμοί (Oracles)' (Comm. cap. 1, pp. 637–638; cp. Diels 32). That is why one finds in the most widely used text-critical edition of the Greek NT, in Nestle and Aland ([27]1993), on the margin of Tit. 1:12 the annotation 'Epimenides, de oraculis'. I look at some of the (partly incompatible) testimonies about Epimenides' life and his works, and I discuss the question whether there is a second quotation from Epimenides in the NT, in my (2013, pp. 125–138).

[55] For two of his exhortations, he chose an extract from the 'Letter to Titus' as his pericope, but neither of them was from the first chapter. On 27.12.1812, the pericope was Tit. 2: 12–15, on 22.11.1818 it was Tit. 3: 1–9.

[56] Church (1946) and Bar-Hillel (1947) raised weighty objections against the paper of which this book is a translation. In Koyré (1947 b) they are simply shrugged off.

F-antinomy is never associated with Epimenides or the 'Letter to Titus' (Spade/Read, § 1.2). But the tradition that Anderson invokes seems to reach back to the renaissance.[57] But whatever its age—we should not follow it. Sometimes a tradition is what Gustav Mahler said about the routine at the Vienna Court Opera: '*Tradition ist Schlamperei* (sloppiness)'.

Back to Russell's exposition of the F-antinomy. Moore rightly criticized also its 'Epimenides' part (Moore 1948/1949, p. 381). When he asks, 'Was this a lie?' Russell suggests—as his next sentence shows ('this contradiction')—that in answering it we will get entangled in a contradiction. Apparently, he takes a liar to be a person whose assertions are always lies, for he moves effortlessly from people who are liars to statements that are lies. (Following Bolzano's footsteps we have criticized this move in Sect. 24.1.1.) So the assertion Russell ascribes to Epimenides really is the assertion that all statements made by a Cretan are lies. Is that the source of a contradiction? If the assertion Russell ascribes to Epimenides is correct then it is a lie—like any other Cretan assertion. But so far, no paradox is in sight, for an assertion can be a lie even if what is asserted is true. (As we saw in Sect. 24.1.2, Bolzano's and Frege's explanations of the concept of lying allow for this possibility.)

Gottlob Frege never discussed the F-antinomy. But some years before Russell, he touched lightly a problem that has an Epimenidean ring, as it were. So lightly that, for all I know, his remark has never been discussed in the Frege literature. It is a one-sentence aside in a manuscript Frege himself never published:

> By your very assertion [you can] contradict what you asserted, in a similar way as a Cretan who said that all Cretans lie. ([Jemand kann] durch seine Behauptung dem, was er behauptet, widersprechen, in ähnlicher Weise wie ein Kreter, der sagte, dass alle Kreter lügen.). (Frege 1969, p. 144, written 1897 or later)[58]

Consider a clear-cut instance of the kind of situation Frege envisages. If a Cretan were to assert that *no Cretan ever makes an assertion* he would contradict what he asserted by his very assertion, more precisely: The fact that he performed this speech-act would falsify its content. Now Frege's Cretan asserts:

(*l*) Whenever a Cretan asserts something he *lies*.

Let us call this Cretan *l*-Epimenides—in order to forestall premature identification with the man quoted in *Tit*. 1:12. Does he also contradict what he asserted by mak-

[57] Jennifer Ashworth examined 47 books printed between 1400 and 1700 that contain discussions of paradoxes, and she found only 9 (the earliest one from 1540) that considered Epimenides' dictum at all. Five of those also examined the F-antinomy, and three even treated the dictum 'as if it were the standard Liar Paradox' (Ashworth 1972, p. 36 and notes 3, 4 u. 13). Actually, there were at least two more 'identifiers' in the period she investigated, and the first one was mentioned in Carl Prantl (1855–1870, vol. 4, p. 170/171). The Florentine humanist Angelo Poliziano (†1494) maintained that Epimenides' dictum and the Pseudómenos come to the same thing (in a letter to the Venetian publisher Aldo Manuzio, in Poliziano (1971, p. 91)). In 1697, Pierre Bayle declared them to be 'le même Sophisme' (414/415). Incidentally, they are also mentioned in one and the same breath in a book Bolzano knew: in Fries ([1]1811, 470=[3]1837, 356).

[58] The final clause is mistranslated in Frege (1979, p. 132): 'that all Cretans are liars'.

ing this assertion? If one lies just in case one asserts what one takes to be false,[59] then asserting (*l*) comes to the same thing as asserting that all Cretans disbelieve whatever they assert. So according to the content of his assertion, *l*-Epimenides is somebody who takes to be false what he is saying, but by performing an act of asserting he presents himself as a person who regards as true what he is saying. If somebody has both the attitude which *l*-Epimenides has according to the content of his assertoric utterance of (*l*) *and* the attitude he gives expression to by his speech-act, then he has blatantly contradictory convictions: He regards the content of his assertion to be false, hence not true, and he regards it as true. Of course, the statement that a person is in such a predicament is itself entirely consistent. But in contrast to the case of a Cretan who asserts that *no Cretan ever asserts anything,* the fact that *l*-Epimenides is a Cretan and utters (*l*) with assertoric force does not falsify the content of his assertion. He could be right in saying what he says even though he disbelieves it. Hence, Frege's 'Cretan who said that all Cretans lie' is far away from the F-antinomy.[60]

Now for the historical question: Did the Cretan *Epimenides* as quoted in the 'Letter to Titus' assert what is expressed by (*l*)? Surely, the answer must be No. But the opposite answer is fairly popular. Quine sometimes reports Epimenides (just as Russell did) as having said that all Cretans are liars, but he also speaks of a 'paradox' that was 'anciently rendered thus: Epimenides the Cretan says that Cretans always *lie*' (the former in Quine 1962, p. 6; the latter in Quine 1953, p. 133; my italics). Alan Ross Anderson first quotes the Greek text and several translations thereof that all render the first conjunct in Epimenides' hexameter correctly, but then he moves from 'are liars' to 'always *lie*', as if these predicates were just stylistic variants of each other (Anderson 1970, pp. 2–3; my italics). Barwise and Etchemendy read 'the claim that Cretans always *lie*, made by the Cretan' into *Tit.* 1: 12 (Barwise and Etchemendy 1987, p. 192; my italics). In his 2003 book *A Brief History of the Paradox,* Roy Sorensen contends that 'Epimenides' remark "The Cretans always *lie*" was quoted for centuries' (op. cit. 94, my italics). If that were so, then Epimenides' remark was *mis*quoted for centuries, but the misquotations are fairly recent.

Let us replace *l*-Epimenides by another Cretan, *f*-Epimenides, who asserted what is expressed by:

(*f*) Whatever is asserted by a Cretan is *false,*[61]

[59] Recall (Df. L) Frege in Sect. 24.1.2.

[60] I do not mean to suggest that nothing paradoxical is in play. We are here in the neighbourhood of what is commonly referred to as Moore's Paradox: 'to say such as thing as "I went to the pictures last Tuesday, but I don't believe I did" is a perfectly absurd thing to say, although what is asserted is something which is perfectly possible logically' (Moore 1942, p. 543). Eldridge-Smith (2004, p. 77) has recognized the kinship of the Cretan who asserts [l], and the 'Moorean' speaker.

[61] Surely, Frege did not use the sentence (l) for saying that Cretans only assert what is false, for he clearly distinguished the concepts lie and false assertion. By contrast, Koyré claims that 'All Cretans lie always' amounts to the same thing as 'all judgements or all assertions made by Cretans are false' (1946, p. 348 ≅ 1947a, p. 10). But 'all judgements … made by Cretans are false' does not entail 'All Cretans lie always': somebody whose judgements are invariably mistaken need not make any assertions, and if he does he might always be sincere.

in symbols ('*C*' for 'is asserted by a Cretan' and '*F*' for 'is false'):

(f) $\quad \forall x\,(Cx \rightarrow Fx)$

The (semi-formal) argument that I shall now construct begins with two assumptions: The assumption that a Cretan asserts that Cretans always assert falsehoods, (1), and the auxiliary assumption that he is right in saying so, (2). In line (3) the argument proceeds in accordance with the *Rule of T-Elimination* that allows us to move from a sentence of the form 'The proposition that *p* is true', briefly: '*T*[*p*]', to the embedded sentence '*p*'.[62] The validity of this rule is ensured by the sense of 'true'. The other rules are natural-deduction rules of classical logic.

1	(1)	$C[\forall x\,(Cx \rightarrow Fx)]$	A[ssumption]
2	(2)	$T[\forall x\,(Cx \rightarrow Fx)]$	A (for →Introduction)
2	(3)	$\forall x\,(Cx \rightarrow Fx)$	(2); T-Elimination
2	(4)	$C[\forall x\,(Cx \rightarrow Fx)] \rightarrow F[\forall x\,(Cx \rightarrow Fx)]$	(3); ∀-Elimination
1, 2	(5)	$F[\forall x\,(Cx \rightarrow Fx)]$	(1), (4); Modus Ponens
1	(6)	$T[\forall x\,(Cx \rightarrow Fx)] \rightarrow F[\forall x\,(Cx \rightarrow Fx)]$	(1), (5); →Introduction

So we can derive from (1) the conditional: if what *f*-Epimenides said is true, then it is false. But we cannot derive from (1) the reverse conditional: if what *f*-Epimenides said is false, then it is true. So, unlike (F_{refl}) 'I am saying something false by uttering this (⇄) sentence', assumption (1) does not engender an antinomy.

The sense of 'is false' is such that an instance of '*F*[*p*]' can always be replaced *salva veritate* by the corresponding instance of '¬*T*[*p*]'. So let us make this substitution in (6) and then apply the argument form 'If *P* then not-*P*; *ergo* not-*P*', (weak) *Consequentia Mirabilis*:[63]

1	(7)	$T[\forall x\,(Cx \rightarrow Fx)] \rightarrow \neg T[\forall x\,(Cx \rightarrow Fx)]$	(6); 'F'/'¬T' Substitution
1	(8)	$\neg T[\forall x\,(Cx \rightarrow Fx)]$	(7); Cons. Mir.

Now perhaps some propositions fall into the truth-value gap, but the proposition which *f*-Epimenides put forward as true, sc. [3] alias [*f*], is certainly not one of them. (Are we not all convinced that it is false?)

9	(9)	$T[\forall x\,(Cx \rightarrow Fx)] \vee F[\forall x\,(Cx \rightarrow Fx)]$	A

We apply the argument form of disjunctive syllogism to (8) and (9) and then we get rid of assumption (1) by conditionalization:

[62] Bolzano endorses this rule (as well as the 'converse' Rule of T-Introduction) in WL IV 114.

[63] In this weak version the Consequentia Mirabilis is valid both in classical and intuitionist logic. On this argument form that received its pretty name in the seventeenth century cp. Bellissima and Pagli (1995).

1, 9	(10)	$\mathrm{F}\,[\forall x\,(\mathrm{C}x \rightarrow \mathrm{F}x)]$	(8), (9); Disj. Syll.
9	(11)	$\mathrm{C}\,[\forall x\,(\mathrm{C}x \rightarrow \mathrm{F}x)] \rightarrow \mathrm{F}[\forall x\,(\mathrm{C}x \rightarrow \mathrm{F}x)]$	(1), (10); →*Introduction*.

There is something about this result that is *beyond belief*. Take any '*p*' you like: if it is false that *p* then not-*p*. Hence, [10] entails the *negation* of [3]. Now the negation of [3] is logically equivalent with the proposition that at least one proposition expressed by a Cretan utterance is not false,—in symbols: $\neg\forall x\,(\mathrm{C}x \rightarrow \mathrm{F}x) -\!|\!|- \exists x\,(\mathrm{C}x\,\&\,\neg\mathrm{F}x)$. Now a truth expressed by an instance of '$\mathrm{C}x\,\&\,\neg\mathrm{F}x$' cannot be identical with the only proposition put forward as true by a Cretan that we know so far, in other words: It must be different from [3], for we saw in line (10) that this proposition is false if it (has a truth-value and) is upheld by a Cretan. So if *f*-Epimenides asserted [3] then the falsity of what he asserts seems to guarantee that there is at least one further utterance made by a Cretan the content of which is *not* false. Now undoubtedly there *is* such an assertion. But can one really deduce this historical truth from *f*-Epimenides' having said what he said? Is no examination of anything else said by a Cretan required? Alonzo Church was right when he declared this result to be *paradoxical*. In a review he observed:

> [M. Koyré says (Koyré 1946, pp. 347–349 ≡; 1947, pp. 9–11)] that Epimenides' statement is self-destructive, in the sense that its truth is disproved by the given fact that Epimenides made it. Apparently M. Koyré is untroubled by a consequence which, though no outright antinomy, might well be classed as paradox. Namely, without factual information about *other* statements made by Cretans, it has been proved by pure logic (so it seems) that *some other* statement by a Cretan, not the famous statement of Epimenides, must have been true. (Church 1946, p. 131)[64]

Note that Church is using here Koyré's labelling: what he misleadingly calls 'the famous statement of Epimenides' is the statement made by our *f*-Epimenides. Now the proposition that a Cretan asserts that whatever is asserted by a Cretan is false is not a claimant to the title 'obviously true'. So we do not have a paradox in the sense I specified in Sect. 24.1.3 unless we read assumption (1) as an assertion of a possibility. For want of a better name, I call the resulting paradox the *Cretan Paradox.*

This paradox can be transformed into an 'outright antinomy' if we add a further assumption that is suggested by a component of Russell's exposition which I have hitherto neglected: 'Epimenides the Cretan said that all Cretans were liars, *and all other statements made by Cretans were certainly lies*.' Quine has improved on this:

[64] (Already A. M. MacIver (1939, p. 65) registered this result as 'very odd'.) Church's hint was first taken up by Arthur Prior in (1958, p. 70 u., 1961, p. 16), then by Bas van Fraassen (1968, p. 150). John Mackie (1973, p. 276) expressed his astonishment in the following way: 'we seem to have discovered a logically necessary connection between two distinct occurrences… This would violate Hume's principle, which in itself is utterly convincing…that there cannot be a logically necessary connection between distinct events.' (Cp. Mates 1981, p. 17; Sorensen 2003, p. 94; Eldridge-Smith 2004, p. 76).

'All *other* statements made by Cretans were indeed false' (Quine 1953, p. 133). So we now assume that every proposition that a Cretan puts forward as true and that differs from what *f*- Epimenides said is *false*, in symbols:

12 (12) $\forall y\,((\mathrm{C}y\,\&\,\mathrm{y} \neq [\forall x\,(Cx \rightarrow \mathrm{F}x)]) \rightarrow \mathrm{F}y)$ A (Russell-Quine)

As the argument from (1) to (6) has shown, under the assumption (1) we can maintain: if the proposition put forward as true by *f*-Epimenides, sc. $[\forall x\,(Cx \rightarrow \mathrm{F}x)]$, is *true* then it is false. On the other hand, if that proposition is false then its negation is true, and that implies that at least one proposition asserted by a Cretan is not false. Now according to the assumption (Russell-Quine), every proposition that is different from the one *f*-Epimenides asserts is *false*. So under the assumption (12), we can maintain: if $[\forall x\,(Cx \rightarrow \mathrm{F}x)]$ is *false* then it is true. Hence, if *both* (1) and (12) are true then the proposition put forward as true by *f*-Epimenides is true just in case it is false (Quine, loc. cit.).[65] Presumably Church would call this 'an outright antinomy'. It is an antinomy in the sense I specified in Sect. 24.1.3 if we read assumptions (1) and (12) as assertions of possibilities. For want of a better name I call the resulting antinomy the *Cretan Antinomy.* It belongs to the family of F-antinomies.

Now for the historical question: Did the Cretan Epimenides according to the 'Letter to Titus' assert what is expressed by (*f*)? Surely, the answer must be No. Nevertheless, Bas van Fraassen maintains: 'Epimenides the Cretan is reported to have said that all statements by Cretans are *false*' (Van Fraassen 1968, p. 150, my italics; similarly Eldridge-Smith 2004, p. 76). The 'report' he refers to is *Tit.* 1: 12–13a, and the paradox he claims to discern there is the Cretan Paradox. Anderson, after having quoted the Greek text, goes on to treat the predicates 'are liars', 'always lie' and 'always utter *falsehoods*', as if they could equally serve to render the first conjunct of Epimenides' dictum (Anderson 1970, pp. 2–3; my italics).

What about the thing Epimenides *did* say, what about the first sentence in his hexameter? Does *that* give rise to the Cretan Paradox or even to the Cretan Antinomy? Russell says about the *NT* passage: '[The paradox] is mentioned by St. Paul, who, however, is not interested in its logical aspect but only in its demonstration that the heathen are wicked' (Russell 1959, p. 77). And Quine seconds him when he says about Ps.-Paul: 'it seems that he missed the point of [the paradox]' (Quine 1962, p. 6).[66] But did Ps.-Paul really 'mention' a paradox and regrettably 'miss its point'? 'Surely', Kripke says, 'no one had a keener nose for paradox than Russell' (Kripke 1975, p. 692, repr. 55). Who is to deny this? But then, there are such things as olfactory hallucinations. Often, physicians tell us, they are caused by traumatic

[65] Compare the first paragraph in Kripke (1975): Since he believes that one can draw the embarrassing conclusion from what is said in the 'Letter to Titus', he beclouds his representation of the paradox by quoting Epimenides rather than f-Epimenides.

[66] Similarly, William Kneale: 'St. Paul does indeed refer to the Epimenides version [of the Liar Paradox], but apparently without realizing that it is a paradox; for he writes …' (Kneale and Kneale, p. 228). Eldridge-Smith (2004, p. 78) also shares this suspicion.

experiences, and as we all know, at the beginning of the last century, Russell did suffer from a logically traumatic experience (Cp. Russell 1967, pp. 151–153).

One might suspect the author of the 'Letter to Titus' of being deaf to a subtle logico-philosophical point, but a century later the philosophically well-versed theologian Clement of Alexandria, too, did not find any logical riddle in that passage. Instead, he applauded it as a confirmation of his own practice of seeking support for some of his views in pagan sources (Clement, op. cit., I, xiv, 59.3):[67]

> Don't you see that he [i.e. the author of *Tit.*] concedes that a Greek prophet has seen a part of the truth and that he has no scruples to use words of a Greek poet when he wants to edify somebody whom he is addressing or to make him feel ashamed?

Another two centuries later, Jerome uses the Epimenides passage for the very same purpose when he replies to a fundamentalist critic:[68]

> Towards the end of your letter you ask me why I occasionally give examples from worldly literature (*saecularium litterarum exempla*) in my writings, polluting the splendour of the Church by the dirt of the heathen (*Ethnicorum sordibus*) … But even Paul the apostle quoted a verse of the poet Epimenides when he wrote to Titus: '*Cretenses semper mendaces, malae bestiae, ventres pigri*.' (Callimachus had also used the first part of this verse.) It is not surprising that the metre is not preserved in a literal translation into Latin.

Jerome quotes the dictum from his own translation of the Scripture that later came to earn the name '*Vulgata* (The Commonly Used [Bible Translation])'. His allusion in brackets (and in *Comm.* 1: 676–684) is to Callimachus' *Hymn to Zeus*:

[6] O Zeus, some say that you were born on the mountains of the Ida,[69]
[7] in Arcadia,[70] say others. Which of them lied (*ἐψεύσαντο*), O Father?
[8] 'Cretans are always liars,' for even a tomb for you, O Lord,
[9] the Cretans built. But you did not die, for you are (live) for ever.[71]

In order to reject the Cretans' claim that Zeus was born on their island,[72] the Hellenistic poet (*fl.* 280 BC) quotes Epimenides' derogatory judgement on his fellow countrymen, and he presents evidence for that verdict: The Cretans say about an im-

[67] By his lights, the study of Greek philosophy, and especially of Plato's philosophy, can serve as a 'propaedeutics' to the acquisition of Christian faith (I, v, 28.1). On Clement's life and work see von Campenhausen (1955, pp. 32–42).

[68] Hieronymus, Epistola LXX, Ad Magnum oratorem urbis Romae (not earlier than AD 399), Sect. 24.2. On the most learned of all Latin 'Church Fathers' see von Campenhausen (1960, pp. 109–150). On quotations, and alleged quotations, from Greek poets in the NT see Künne (2013, pp. 131–138).

[69] In Crete.

[70] At the centre of the Peloponnese.

[71] Callimachus of Cyrene, poet, scholar and librarian in Alexandria, in: Kallimachos (2004, p. 388) (lines [8]-[9] also in *Diels* 32). What Callimachus took to be a blasphemy on the side of the Cretans, the 'Church Fathers' welcomed as a sign of disillusionment concerning the Olympian Gods. See below, *Appendix* III.

[72] Not only the Cretans believed this (cp. Hesiod, Theogony, lines 474–484).

mortal God that he is buried.[73] Of course, two or three centuries earlier, Epimenides himself may have had a completely different reason for his indictment of the Cretans. In the remainder of his letter, Jerome tries to convince the recipient in Rome that the great Alexandrian theologians Clement and Origen benefited a lot from their knowledge of secular literature.[74]

Roy Sorensen wonders at 'thousand years of Christian incomprehension of the Liar Paradox', for 'one might expect it to be kept steadily before the Christian eye because it is repeated in the Bible' (Sorensen 2003, p. 199, 197). Well, perhaps the sight of the 'Letter to Titus' could not start the Christians off because the F-antinomy is not at all 'repeated' there…

Since the middle of the twelfth century, the F-antinomy was discussed again and again, and now it is Paul Vincent Spade's and Stephen Read's turn to express amazement: 'not a single medieval author is known to have discussed or even acknowledged the logical and semantic problems this text [sc. *Tit* 1: 12–13a] poses' (Spade/Read, § 1.2). Thus, when Aquinas comments on the 'Letter to Titus', he is only concerned with what Clement and Jerome had emphasized many centuries earlier:[75]

> [St. Paul] confirms here the testimony of one of their poets, namely that of Epimenides… Gloss. Here we can see that he acknowledges a testimony for a truth wherever he found it (*accipit testimonium veritatis ubicumque invenerit*). That's why the apostle quotes at several places sayings of the heathen (*dicta gentilium*).

The silence of the 'Church Fathers'[76] and of medieval philosophers is quite easily explained: The Epimenides passage does not pose any semantic or logical problems at all. Rather, such problems were imposed on the passage by the twentieth century philosophers greedy for paradox.

Apparently, the only analytical philosopher who ever bothered to have a close look at Epimenides' notorious dictum was G. E. Moore. He even paid attention to the gender of the first word in '*Κρῆτες ἀεὶ ψεῦσται*'. It is the plural of a *masculine* noun, '*ὁ Κρής*'. The women of Creta are called *Κρῆσσαι* (as in the title of a lost play by Aeschylus). So perhaps Moore is right: Epimenides only wants to rebuke the *male* Cretans.[77] As for the last word, nobody doubts that '*ψεύστης*' really means

[73] This is a manifestation of their mendacity only if they believe that Zeus is not buried. Actually, in line [7] the wording of the question is not as unequivocal as my translation suggests: Isn't the question rather which of the competing claims is true? But then, it is not reasonable to insist on conceptual precision in a hymn…

[74] In Spade (1973, pp. 296–297), one can find many further pieces of evidence for this way of treating the reference to Epimenides' dictum in the Patrologia Latina. I quote, and comment upon, a passage in Augustine in my (2013, p. 60 f.).

[75] *Super Epistolam S. Pauli ad Titum lectura* (a lecture Aquinas presumably gave between 1265 and 1267 in the Dominican Convent in Rome), towards the end of lectio 3.

[76] This somewhat tendentious label is customatily used for Greek, Roman and Syrian writers of Christian confession who lived between 100 and 500.

[77] Of course, whether Moore is right about this is entirely irrelevant for the question whether the dictum gives rise to a paradox. But if he is right, the women of Creta are saved from the discrimination. In the middle and long run, at least, even the male Cretans did not resent Epimenides for

liar.[78] But what about the second word—how is 'ἀεί (always)' in this context to be understood? English sentences of the form '*A*s are always *B*' are sometimes just stylistic variants of '*A*s are without exception (invariably) *B*' or 'All *A*s are *B*'. For example, the word 'always' in 'Even numbers are always divisible by 2 without remainder' is certainly not a *temporal* adverb.[79] Does this also hold of the Greek adverb 'ἀεί'? No, answers Moore: 'I do not think "ἀεί" is ever used in this particular way, in which "always" is sometimes used in English' (Moore 1948/1949, p. 378).[80] Consequently, he rejects Russell's rendering of Epimenides's saying as 'All Cretans are liars'. Moore offers the following paraphrase of the Greek sentence: 'If you ever meet a male Cretan, you will rarely be wrong in assuming that he is an habitual liar.' He is certainly right in thinking that no logician has a reason to get excited when reflecting on *this* sentence as uttered by a Cretan. The authors of the *Port-Royal Logic* anticipated the Moore's deflationary reading of Epimenides' dictum when they assimilated it to statements like 'All old people praise the past' (Arnauld and Nicole [6]1685, II. Part, Chap. xiii):

> We call universality *metaphysical* if it is perfect and without exceptions, such as 'Every man is a living being' … And we call universality *moral* if it admits exceptions, for in moral matters [i.e. in matters of *mores*, of manners and customs] we content ourselves if things are usually thus and so, *ut plurimum,* as when St. Paul quotes and confirms *Cretes semper mendaces, malae bestiae, ventres pigri* … for some inhabitants of the island might not have had the vices that were common to the others.

But I am afraid that Moore's hypothesis about the use of the Greek word 'ἀεί' is definitely mistaken. Here are three pieces of evidence for a nontemporal use of this word in the *Corpus Aristotelicum*: (1) When Aristotle says, 'Necessarily a negation always (ἀεί) says something true or false', he means that necessarily *every* negation is bivalent. (2) When he says, 'In this figure the conclusion is always (ἀεί) particular', he means that *every* argument in a certain figure has a particular conclusion. (3) And when he says, 'What follows from a truth is always (ἀεί) true', he means that

---

the defamation. In Plato's Nomoi (I, 642 D-E), the Crete Kleinias praises Epimenides as a 'divine man'. Nowadays a street in the capital Heraklion bears his name (Mates 1981, p. 164 and Google Maps), and so does the cultural association of Panormo at the northern coast of the island (as you can check on the Internet).

[78] The earliest passages LSJ adduces for this usage are from the Iliad and from Herodotus.

[79] When Russell (1905, p. 42; 1910, p. 16; 1918, p. 230) suggests reading '∀x *Fx*' as '*Fx* is always true', he certainly does not want us to understand the adverb as temporal. His paraphrase either means that each substitution instance of the open sentence '*Fx*' is true (expresses a truth), which would make Russell an advocate of the substitutional interpretation of the quantifiers, or it means that each object complies with the condition that is signified by the open sentence.

[80] Moore's conjectures about such issues deserve to be taken seriously. In his autobiographical notes he recalls: '[In my final years at school] almost all my time was spent on Greek and Latin', and '[in 1892] I came up to Cambridge expecting to do nothing but Classics there, and expecting also that afterwards, all my life long, my work would consist in teaching Classics to the Sixth Form of some Public School—a prospect to which I looked forward with pleasure.' As an old man he still has fond memories of Henry Jackson's lectures on Plato and Aristotle he heard when his friend Russell had persuaded him to study philosophy (Moore 1942, p. 5, 13, 20. Cp. also Moore 1962, pp. 365–369 on Plato, Nomoi X, 895 D).

everything entailed by true premises is true.[81] So we need not have any philological scruples when we report Epimenides as saying that all Cretans are liars.

Do we run into logical trouble if we insist (*pace* Moore and his French predecessors) on the non-deflationary reading of this generalization? I do not think so. The predicates in Epimenides' hexameter prevent any such trouble. To say now of a male contemporary that he is a notorious *liar* is not to say what he is doing right now, any more than to say of him that he is an *evil beast* and a *lazy glutton* is to say which activities he is currently engaged in. If you make the former accusation you ascribe mendacity to that man, which is a habit rather than an act. Epimenides may very well be absolutely right when he ascribes this reprehensible habit to himself and to all his (male) compatriots, and what Ps.-Paul asserts is that his witness *is* indeed right. *There is nothing paradoxical about this,* even if both speakers intend strict, exceptionless universality. Quine actually conceded this, without adapting his terminology to this observation and without retracting his censure on Ps.-Paul: 'The ancient paradox of Epimenides the Cretan, who said that all Cretans are liars… is untidy; there are loopholes. Perhaps…Epimenides was a liar who occasionally told the truth' (Quine 1962, p. 6). A paradox with loopholes is not a paradox, any more than a triangle with four sides is a triangle. Why not acknowledge that there is no paradox here, which implies that the author of the 'Letter to Titus' cannot have 'missed the point of it' (as Quine maintained a few lines earlier)? All in all, Jerome understood the Epimenides passage far better than most logicians who refer to it: 'Just because the Cretans have the vice of being mendacious (*vitium mendacii*) we need not assume that they never said something true (*non statim et verum non aliquando dixerunt*)' (Hieronymus, *Comm.* I: 723–726).[82] As you will recall, this was also the message of Bolzano's analysis of the fallacious argument concerning self-confessed liars in *WL III*. What the author of the 'Letter to Titus' maintains (even under the non-deflationary reading) does not give rise to the Cretan Paradox, let alone to the Cretan Antinomy.[83]

24.2 'A Self-Destructive Sentence'?

24.2.1 Fra Girolamo on the Antinomy of Falsity

In a long appendix to § 19 of the first volume of his *Wissenschaftslehre,* Bolzano discusses the exposition, and the diagnosis, of the F-antinomy in Girolamo Savonarola's *Compendium logicae* (Bolzano, *WL I* 78–80; Savonarola, *CL* 151,

[81] (1) Int 10: 20a34, cp. Int. 2: 16a32–16b5, (2) An. Pr. II.7: 58b41–59a1; (3) An. Post. I.6: 75a5–6. I owe the references under (1) to Hermann Weidemann.

[82] As we will see in Appendix IV, it was only a line in the Book of Psalms that made Jerome consider (if only in passing) an argument that sustains a paradox.

[83] It redounds to the honour of New Testament scholars that they hardly ever projected a logical problem on Titus 1: 12–13a. I inveigh against an inglorious (British) exception in my (2013, pp. 144–147).

lines 6–24).[84] The author of this manual is better known as the irate Dominican preacher of repentance and reform who, at the instigation of the Roman Curia, was tortured, hanged and then (to be on the safe side?) burnt at the stake in Florence in 1498. The Florentine humanist Pietro Crinito testifies to the high esteem in which Savonarola's learned contemporaries held his philosophical competence. He recalls a conversation about ancient philosophy between the Friar and his admirer Giovanni Pico della Mirandola: 'Recently we sat together *in Marciana academia* [i.e. in the Convent of San Marco in Florence] with Hieronymus Savonarola who stands out in our time in almost every area of philosophy (*qui aetate nostra in omni prope philosophia praestat*)' (Crinito 1504, lib. III, cap. ii).[85] Fra Girolamo wrote his *Compendium logicae*—or a first version of it—in the Convent of San Marco around 1484. It was to serve as a textbook for the novices. At the monastic schools of the Dominican order, a *studium logicae* of 2 years was obligatory. For a few decades, Savonarola's manual was widely used in those schools, in 1516 and 1596 it was printed even in Saxony, but then it soon fell into oblivion—like all his philosophical work.[86]

In his exposition of the F-antinomy and in his attempt at defusing it, the Frate uses the following *F*-sentence:

> *This is false*—under the assumption that the subject term is used to point at that very sentence *(hoc est falsum, posito quod per subiectum demonstretur ipsamet propositio)*. (*Compendium logicae* 151: 6–8, also quoted in *WL*)

Let us call a sentence *S* self-referential if, and only if, it contains a singular term that denotes *S*. By this standard, '*Hoc est falsum'* is self-referential (under the intended reading). Savonarola offers a second example that is not mentioned by Bolzano. It also gives rise to an antinomy though only under a counterfactual assumption:

> *Every sentence is false*—under the assumption that this is the only sentence in the world (*omnis propositio est falsa*, *si ponatur quod solum sit in mundo ista proposition.* (*CL* 150:23–151, 4[87])

[84] Fra Girolamo examines the antinomy in Sect. 24.18 of the *Liber decimus: De syllogismo sophistico* (we would call the libri of this slim manual chapters) under the heading '*Insolubile propositum nec est concedendum nec negandum* (if an insoluble is put forward it should be neither accepted nor denied)'. I shall explain the terminology below.

[85] Pasquale Villari (1859, p. 99) refers to this passage as well as to the letters by Ficino and Poliziano who also describe the friar as a man of distinguished learning.

[86] Bolzano used the 1516 edition. In Appendix V you find a report on the early history of the book and its sparse reception in the last two centuries as well as some pointers to Savonarola's other philosophical writings.

[87] A century earlier, Jean Buridan had employed the same sentence for his Seventh *Sophisma,* but he made a different counterfactual assumption: The sentence is uttered in a possible world in which there *are* other utterances as well but only false ones [Johannes Buridanus, *Sophismata*, cap. 8, *De propositionibus habentibus supra seipsas reflexionem: Septimum sophisma est hoc vocatum insolubile,* in: Buridanus 1977, pp. 133–137; 1982, pp. 60–73]. The antinomy thereby engendered is a variant of what I dubbed the *Cretan Antinomy* in Sect. 24.1.4—Medieval philosophers used the word 'sophisma' for logically enigmatic sentences—it does not have the pejorative connotations that it has for some ancient and all modern authors (cp. note 24, note 57 above).

This sentence does not contain a singular term that denotes it, hence it is not self-referential in the sense I specified—notwithstanding the fact that in a possible world in which it is the only sentence (as well as in others possible worlds in which it is not quite as lonely) it says something *about* itself, as it were. Savonarola's second example suggests that by '*propositio*' he means perceptible *occurrences* of sentences—*token sentences*. ('Sentence', here as elsewhere in this chapter, is short for 'declarative sentence'.) As a matter of fact, he is thinking of *oral* occurrences, for at the beginning of his manual he points out that 'arguments consist of sentences', that 'sentences consist of simple words (*propositiones ex simplicibus vocibus componuntur*)' and that a word (*vox*) is a special kind of noise (*sonus*) (*CL* 4–5). If one can hear the components of sentences then sentences must also be acoustically perceptible. So by '*propositio*' the Frate means utterances, not propositions. What we call proposition, authors of the Middle Ages and the Renaissance used to call *dictum propositionis* (what is said by an utterance) or *enuntiabile* (assertible). Different utterances of the type sentence '*Hoc est falsum*', self-referentially understood, express different propositions, since they are about different utterances.

Bolzano translates Savonarola's three-word sentence loosely and wordily as '*Die Rede, die ich so eben führe, ist falsch* (the utterance I am making right now is false)', and then he replaces it by formulations that go better with his own conception of truth-value bearers, first by '*Was ich jetzt sage, ist falsch* (What I am saying right now is false)' and then by '*Was ich so eben* [or *jetzt eben*] *behaupte, ist falsch*' (*WL I* 79–80),

(*F*) What I am asserting right now is false.

There is no indication that the illocutionary difference between saying and asserting matters to Bolzano in this context. Clearly, (*F*) is a close relative of the father of all *F*-sentences, i.e. '(*ἐγὼ*) *ψεύδομαι*', '*(ego) mentior*'. The content of an act of assertion, what is asserted, is a proposition or *Satz an sich*. The appendix on the F-antinomy belongs to that section of the *Wissenschaftslehre* that is to clarify 'what the author means by a proposition (*was der Verfasser unter einem Satze an sich verstehe*) (*WL I* 76). For Bolzano (as for Frege), the primary truth-value bearers are neither token sentences nor type sentences. Different utterances of the English sentence '7 is a prime number' (and of the German sentence '*7 ist eine Primzahl*') express one and the same proposition. An indexical sentence like 'I am very busy today' expresses in my mouth on 12th May, 2012 a different proposition than in the mouth of any other speaker or on any other day, and some of those propositions are true, some are false (Cp. *WL I* 113; *II* 77–78). All utterances of 'Today I am older than yesterday' express true propositions, but if the speaker, or the day of the utterance, are not the same, then different truths are expressed because the denotation of the indexicals varies. Hence, there is no such thing as *the* proposition expressed by utterances of (*F*).

In the context of a discussion of the F-antinomy, (*F*) is not optimal. For one thing, Moore's objection can be raised again. Suppose Mr. X who no longer loves Mrs. Y tells her in a letter that he still loves her while self-critically commenting on

his letter by uttering (*F*). Then no paradox arises. Furthermore, if one wants to transform Savonarola's *F*-sentence into a sentence that no longer treats linguistic entities as a truth-value bearers, there is no need to invoke personal or temporal indexicals. Without provoking Moore's objection, we can use:

(ψ) The proposition that is expressed by this ($\rightleftarrows$) sentence is false.[88]

Here, the number of indexical elements is reduced to one (as in Savonarola's paradigm). In order to avoid a multiplication of truth candidates that is irrelevant for the antinomy, I beg the reader to regard tokens of the demonstrative definite description 'this sentence' that are to be understood reflexively always as denoting the *type sentence* in a token of which it occurs. Thus understood, all utterances of sentence (ψ) express the *same* proposition (if they express a proposition at all). I shall use (ψ) as 'propositionalist' counterpart to Savonarola's *F*-sentence '*Hoc* ($\rightleftarrows$) *est falsum'*.

The Frate classifies '*Hoc est falsum'* under the intended reading (as well as '*omnis propositio est falsa*' in a possible world in which no other sentential utterance occurs) as an *insolubile*.[89] In the Middle Ages, this somewhat defeatist title was widely used for paradox-inducing sentences.[90] Savonarola explains it as follows:

> An Insoluble (*insolubile*) is a sentence that destroys itself (*propositio seipsam destruens*). For if it is said to be true it follows that it is false, and if it said to be false it follows that it is true (*Quia si dicitur quod est vera sequitur quod sit falsa, et si dicitur quod est falsa sequitur quod est vera*). (*CL* 150:21–23)

This also holds of the proposition that seems to be expressed by (ψ). If a sentence is used for ascribing the property *X* to an object *a,* the proposition expressed by that sentence is true iff *a* really has *X* (cp. *WL I* 124). Suppose utterances of sentence (ψ) express a proposition, call it '[ψ]'. Then in uttering (ψ), one ascribes the property of being false to [ψ]. If the proposition expressed by that utterance is true, then the object referred, that is, [ψ], *has* the property that is ascribed to it in the utterance, namely falsity. In a nutshell:

(I) If [ψ] is true then [ψ] is false.

On the other hand, if the proposition expressed in an utterance of (ψ) is false, then it has the very property that is ascribed to it by that utterance. If it has the property that is ascribed to it then it is *true*. In fine:

---

[88] You find the same formulation in Moore (1962, p. 292, 314). (The psi is meant to prompt memories of the ancient precursor.)

[89] Not 'insolubilium', as Rescher has it: (2001, pp. 195–98).

[90] 'On Insolubles (De insolubilibus)' was the standard title for treatises on such sentences. Cicero's discussion of the F-antinomy that we examined in Sect. 24.1.3 was available in the MA, but apparently it had no impact on those treatises, for otherwise their authors would presumably have used Cicero's label 'inexplicabilia' for the paradoxes, which none of them does (cp. Spade 1973, II. pp. 299–300 with reference to De Rijk; Spade/Read § 1.1).

(II) If [ψ] is false then [ψ] is true.

For Savonarola as for Bolzano, it is a matter of course that nothing is both true and false:[91] What is false is not true, and what is true is not false. Let us call this the *principle of exclusiveness,* '(Excl)' for short. If we apply the argument form of hypothetical syllogism twice, we obtain:

(1)	If [ψ] is true then [ψ] is not true	(I), (Excl); *Hyp. Syll.*
(2)	If [ψ] is false then [ψ] is not false.	(II), (Excl); *Hyp. Syll.*

Now, we invoke twice the argument form 'If *P,* then not-*P*; *ergo* not-*P*', the weak version of *Consequentia Mirabilis* (Cp. Sainsbury 2009, p. 128):

(3)	[ψ] is not true	(1); CM (weak)
(4)	[ψ] is not false	(2); CM (weak).

From (3) and (4), we can derive the conclusion that the proposition [ψ] falls into the gap between the two classical truth-values:

(Gap)	[ψ] is neither true nor false.	(3), (4); &-Introduction.

If the assumptions and the rules that this argument relies upon are true and correct, respectively, then there is a least one truth candidate that is neither true nor false. For philosophers like Savonarola and Bolzano who take it to be obvious that no truth candidate falls into the gap, this result is *beyond belief* (paradoxical). The Florentine monk accepts bivalence for sentences, the Prague priest endorses it for propositions (Cp. *WL II* 33). From the above result, Savonarola draws the conclusion that we must give up the assumption that an utterance of 'This ($\rightleftarrows$) is false' is a *propositio* at all. He writes:

> If it is objected that every sentence is either true of false, then one should reply that these are not sentences. For they do not comply with the definition of a sentence, namely that it is a true or false utterance (*oratio*). (*CL* 151:19–22, also quoted in *WL*)

If Savonarola were to use the concept of a proposition he would agree with Moore: we must give up the assumption that (ψ) expresses a proposition (Moore 1948/1949, pp. 383–384, cp. 1962, pp. 171–172),[92] in other words, '[ψ]' is as empty a singular term as 'the golden mountain' and 'the largest prime number'.

The Frate goes on to emphasize a semantic feature of the nominal phrase 'self-destructive sentence' that distinguishes it from syntactically similar constructions like 'self-contradictory sentence' or 'short sentence'—just as it distinguishes 'would-be virtuoso' from 'Russian virtuoso'. While every Russian virtuoso is a vir-

[91] Only self-styled 'Dialetheists' like Priest (and the main speaker in the small dialogue by Michael Eph.) deny this: see Sect. 24.1.3.

[92] This is also the strategy of which William Kneale (1971, esp. pp. 241–43 and 1972, esp. pp. 321, 330–31) and Jordan Howard Sobel (1992) believe that it can block the F-antinomy.

tuoso, no would-be virtuoso is a virtuoso, and while all self-contradictory, or short, sentences are sentences, no self-destructive sentence is a sentence.[93] Savonarola uses an example of Aristotelian provenance:

> Admittedly, [such sequences of words] have the shape of a sentence. But just as a dead man has the shape and the appearance of a man without being a man,[94] so [such sequences of words] are called self-destructive or insoluble sentences without really being sentences. (*CL* 151:22–26; also quoted in *WL*. Cp. *CL* 198:19)

In Shakespeare's *Hamlet,* the gravedigger answers the prince's questions in an Aristotelian spirit (V.1): 'What man does thou dig it for? *For no man, sir.* What woman, then? *For none, neither.* Who is to be buried in ´t? *One that was a woman, sir; but, rest her soul, she's dead.'* So according to the Frate, *F*-sentences are not really sentences although they look like sentences, just as dead women are not really women although for a while they look like women.

In the Middle Ages, philosophers who claim that *F*-sentences are not sentences (do not express propositions) were called *cassantes,* nullifiers: they 'claim that somebody who says that he is saying something false does not say anything (*Cassantes … dicunt quod dicens se dicere falsum nihil dicit*)'.[95] If we were to allow ourselves to use indirect speech as liberally as those nullifiers did, formulations like this would be permitted: 'If you say that the slithy toves did gyre and gimble in the wabe, you do not say anything'. Clarity would be served by semantic ascent: 'If you utter the words "the slithy toves did gyre and gimble in the wabe" you do not say anything.' Similarly here, the nullifiers maintain that somebody who utters the words 'I am saying something false' does not say anything. As we saw in Sect. 24.1.3, Alexander of Aphrodisias was an early *cassans*.

Jan Berg ascribes to Savonarola an argument for nullification that does not invoke the principle of bivalence. According to Berg, the Frate argued: (ψ) is not really a sentence, 'since (ψ) contains itself as logical subject' (Berg 1962, p. 59, repeated in 1985, p. 15).[96] But in the *Compendium,* there is not the slightest trace of such a reasoning. This misinterpretation seems to be triggered by *Bolzano*'s text. He writes:

> One might very well think that Savonarola is right, especially as the subject of a proposition can never be that proposition itself, just as a part can never itself constitute the whole (*Man sollte glauben, daß S*[avonarola] *recht habe, und zwar besonders darum, weil das Subject*

[93] The semantical difference between these two types of 'noun companions' also plays an important role in Bolzano (WL I 92, 121, 138, 257f, WL II 213. Cp. Künne 1997, p. 329; esp. Benjamin Schnieder 2007, pp. 530–537).

[94] Cp. Aristotle, Int. 11: 21a21–23; Meteor. IV.12: 389b31; De part. anim. I.1: 640b34 f. (In Savonarola's personal estate there were copious excerpts from most works of Aristotle—apart from the Organon (Garin 1959, p. 206)).

[95] Quoted after Kneale and Kneale (1962, p. 228; cp. Spade/Read § 2.5). Paolo Veneto presents the position of the *cassantes* as the 5th strategy for defusing the paradoxes, and he rejects it (Bocheński, p. 281). Arianna Betti (2006, p. 71) has pointed out that Jan Łukasiewicz is also a nullifier (1915, pp. 29–30). Arguably, the same holds of Tarski, for according to him natural language truth ascriptions are not well-formed sentences.

[96] I took the liberty of replacing Berg's abbreviation of an F-sentence by my own.

> *eines Satzes doch nie er selbst seyn kann, so wenig, als ein Theil das Ganze ausmachen kann*). (*WL I* 79.)

Apparently, Berg thinks that Bolzano is here recapitulating *Savonarola's* argument for the contention that sequences of words like (ψ) are not sentences.[97] But Bolzano is here playing the *advocatus diaboli* (may Fra Girolamo pardon my formulation!) in order to show that for such a reasoning one cannot legitimately appeal to the undeniable mereological truth that:

(Mer) $\forall x\, \forall y\, (x \text{ is a proper part of } y \rightarrow x \neq y)$

He shows 'with the help of the distinction between a proposition and a concept of that proposition' that this appeal is illegitimate (*WL I* 79):

> Not the proposition itself…but only the concept of this proposition is the subject concept in that proposition (i.e. in [*F*] or [ψ]). That this distinction is well grounded is shown by the fact that not only here but in all cases the object itself is to be distinguished from its concept, if one does not want to get entangled in gross absurdities (*Nicht der Satz selbst… sondern nur die Vorstellung von ihm, macht die Subjectvorstellung in jenem Satze*) *aus. Daß diese Unterscheidung gegründet sey, beweiset der Umstand, daß man nicht etwa nur hier, sondern überall die Sache selbst von dem Begriffe derselben unterscheiden muß, will man sich nicht in die gröbsten Ungereimtheiten verwickeln.* (loc. cit.)

In other words, the concept that is the subject of [ψ] is expressed by the singular term 'the proposition that is expressed by this sentence' if reflexively understood, and [ψ], far from being identical with that (individual) concept, falls under it. The same holds mutatis mutandis for propositions expressed by [*F*].[98] And a sentence *S* need not itself be its own subject in order to be about *S*—it is sufficient if it contains a term that refers to *S*. So Savonarola does well not to appeal to (Mer) when he argues that his *F*-sentence is only a mock sentence. (As we shall see in Sect. 24.3.1, a variant of the bad argument occurs in Wittgenstein's *Tractatus*.)

Fra Girolamo's reasoning is very different: Suppose '*Hoc* (⇄) *est falsum*' is a real sentence. Then one can show that it is neither true nor false. But no real sentence falls into the truth-value gap. So that chain of words is not really a sentence. Recast in the propositionalist style used above, the argument runs as follows: Suppose (ψ) expresses a proposition. Then one can show (as we did above) that this proposition

[97] Larry Hickman (1976, pp. 398–399) contends: 'Savonarola akzeptierte eine Fassung der restrictio-Lösung des Paradoxons, die behauptet, daß ein Teil einer Proposition nicht für die ganze Proposition, deren Teil sie ist, stehen (supponere pro) darf.' There is no trace of this 'solution' in the text of the manual. Though Hickman refers to Savonarola's Compendium logices [sic], he seems not to have read the pertinent passage in that book but only Bolzano's appendix to WL § 19 in the edn. Bolzano (1963, pp. 25–27) and Berg (1962). The editor Kambartel relies on Berg's interpretation, and he also refers to Savonarola's manual by using the Grecized title of its later Venetian editions (Appendix V). Elke Brendel (1992, p. 27, 42, cf. 217) also relies on Berg's interpretation of the Frate. But as we shall see in Sect. 24.3.3, she emphasizes (for the first time, I think) an essential insight in Bolzano's discussion of the F-antinomy.

[98] Admittedly, it is not absurd to invoke (Mer) if one wants to argue that utterances of (ψ) or (F) fail to express *Russellian* propositions. But as was to be expected, Bolzano employs *his* conception of a proposition: the proposition that *a* is thus and so contains the sense of '*a*' (rather than the object *a* itself) as a component. He would agree with Frege's 'Mont Blanc' objection against Russell.

is neither true nor false. But no proposition falls into the truth-value gap. So (ψ) does not express a proposition.

This reasoning invites at least three objections. *Firstly,* one may reasonably wonder whether the principle of bivalence is really beyond doubt.[99] *Secondly,* the argument given above does not *prove* that [ψ], given (I) and (II), falls into the truth-value gap. If we endorse *biv*alence (as Savonarola and Bolzano do) we might as well argue in the following way:

(I)	If [ψ] is true then [ψ] is false.
(II)	If [ψ] is false then [ψ] is true.
(Biv)	A proposition is false iff it is not true.

From these assumptions we can derive:

(1)	If [ψ] is true then [ψ] is not true.	(I), (Biv); *Hyp. Syll.*

Applying the weak version of *Consequentia Mirabilis* we obtain:

(2)	[ψ] is not true.	(1); CM (weak)

From our assumptions we can also derive:

(3)	If [ψ] is not true then [ψ] is true.	(II), (Biv); *Hyp. Syll.*

Now we apply the strong version of *Consequentia Mirabilis,*[100] i.e. the argument form 'If not-*P* then *P*; *ergo P*', and obtain:

(4)	[ψ] is true.	(3); CM (strong).

[99] In the second half of the twentieth century several logicians regarded giving up the principle of bivalence even as part of the *solution* of the F-antinomy. This is the least common denominator of Martin (1967), van Fraassen (1968) and Kripke (1975). Because of some antinomies lurking in the neighbourhood (such as the antinomy van Fraassen called the 'Strengthened Liar'), the rejection of bivalence requires some flanking measures. (The antinomy just mentioned has its source in the sentence *The proposition expressed by this* (⇄) *sentence is not true,* hence it would better be called the Antinomy of Untruth.) In the Middle Ages, philosophers who held that *F*-sentences fall into the truth-value gap were called mediators (*mediantes*). So far, only one mediator has been spotted in the time from 1132 to 1372, an English Benedictine with the somewhat infelicitous name Roger Swyneshed [Spade/Read § 3.2; ed. and comm. in Spade 1988, chs. VII and VIII]. At the end of the MA, Paolo Nicoletti Veneto presented the position of the mediators as the 6th of 15 attempts to solve the paradoxes, and he rejects it: the mediators 'are mistaken, for every *propositio* is true or false, and each *insolubile* is a *propositio*' (Paulus Venetus, *Logica Magna* (written ca. 1396–1399), part II.15, *De Insolubilibus*). Currently (2012), there exists no critical edition of this part of that monumental logic. Some pieces of it are translated in Bocheński (pp. 280–293, here 281).

[100] In his (1810, pp. 124–126) and in WL (IV, 280), Bolzano thematizes the strong version of CM that is valid only in classical logic, and in WL (I, 145, § 31) he applies it in order to show that there is at least one truth (Cp. Stefania Centrone 2012a, b).

Finally, we derive:

⊥	[ψ] is true & [ψ] is not true	(2), (4); &-Introduction.

This seems to show that the *F*-sentence (ψ) is the source of a blatant contradiction if we stick to bivalence and to the rules of classical propositional logic employed in this deduction.

Furthermore, the above trio of assumptions can also be used to derive a result that is the polar opposite of truth-value gap, namely *truth-value agglomeration* ('glom', for short). In order to obtain this result, we continue after (2) as follows:

(3*)	[ψ] is false	(Biv), (2); Modus Ponens
(4*)	[ψ] is true	(II), (3*); Modus Ponens
(Glom)	[ψ] is true & [ψ] is false.	(3*), (4*); &-Introduction

This is the very conclusion that Alexander of Aphrodisias and Michael of Ephesus claimed to be derivable (Sect. 24.1.3). If you accept glom as true (as do the speaker who gets the final word in Michael's postscript and his descendant Graham Priest) you deny that the principle of exclusivity is universally valid—in other words, you contradict all those who share Alexander's conviction that truth and falsity are contrary properties.

Thirdly, the nullifiers' strategy to deny that *F*-sentences really are sentences leads to very implausible consequences. Saul Kripke has pointed out that 'quite ordinary assertions about truth and falsehood are liable, if the empirical facts are extremely unfavourable, to exhibit paradoxical features' (Kripke 1975, p. 691/repr. 54). He shows this with the aid of a pair of utterances (The Watergate example: 691–692/repr. 54–55). I think that a single utterance can also serve as evidence for his contention. Consider the sad case of John. Ever since he read Mark Twain's mockery about 'The Last Words of Great Men',[101] he was afraid that he might make a fool of himself on his deathbed. Being painfully aware of his cognitive shortcomings, he wrote on a sheet of paper:[102]

(Ω)	The proposition expressed in my last words is false.

And then fate was doubly unkind to him. I spare you the details: these *were* John's last words. His sudden death was bad enough, but to crown it all it transformed what he took to be a melancholy prediction into an utterance of an *F*-sentence. Whether an ascription of falsity is self-referential or not cannot always be found out by inspecting it, so *a fortiori* one cannot read off from its intrinsic properties whether it is self-referential in a logically embarrassing way (Kripke 1975, p. 692/repr. 55; Mates 1981,

[101] 'A distinguished man should be as particular about his last words as he is about his last breath … He should never…trust to an intellectual spirit at the last moment to enable him to say something smart with his latest gasp and launch into eternity with grandeur.' (Mark Twain 1869).

[102] With atemporal 'is': John always shunned tenses when he made solemn pronouncements.

p. 24). If he were asked to comment on my grim example, Savonarola would have to say: 'Whether a sequence of words in language *L* that can be used to make true or false assertions really is a sentence as uttered by an L-speaker at time *t* sometimes depends on what happens to that speaker after *t*.' But can the sentencehood of John's utterance depend on the span of his life? This is hardly an attractive position. It does not become more attractive if you replace 'is a sentence' by 'expresses a proposition'. And then, there is a compositionality problem. Suppose my pitiable John had first uttered (Ω) as embedded in one of the frames 'I am afraid that ( )', 'Possibly, ( )' or 'If ( ) then ( )' before he uttered it by itself. Then his earlier utterance would have expressed truths. How can that be the case if his subsequent utterance of (Ω) alone is not a *propositio* at all, and does not express any proposition?[103]

24.2.2 Bolzano Versus Savonarola (I)

Bolzano also rejects nullification:

> Still, I dare profess the contrary opinion, and I believe that common sense, too, will decide in my favour. Which teacher of languages will hesitate to call the words 'What I am saying right now is false' a sentence that expresses a complete sense? (*WL I* 79)

Surely, he would repeat this for

(*F*) What I am asserting right now is false

(ψ) The proposition that is expressed by this ($\rightleftarrows$) sentence is false.[104]

In fact, the conception of propositions spelt out in his book obliges him to take this stance. For this theory includes the claims (1) 'that sentences of the form "*A* is *B*" never have a sense that differs from the sense expressed by "*A* has *b*", if "*b*" represents the *abstractum* that belongs to the *concretum* "*B*"' (*WL II*, 10),[105] and (2) that *each* sentence of the form "*A* has *b*" expresses a proposition 'if "*A*" stands for any concept of an object and "*b*" stands for any concept of a property' (*WL I*, 393).[106] The noun phrases that occupy the '*A*' position in (*F*) and (ψ) express concepts of an object (*Gegenstandsvorstellungen*), and 'falsity', the nominalization of the general term in in the '*B*' position, expresses a concept of a property (*Beschaffenheitsvorstellung*). Even if it is an open question whether the subject term in (*F*) or (ψ) denotes anything, it expresses an objectual concept, for a concept is objectual just

[103] This is one of the objections that Anil Gupta and Nuel Belnap (1993, pp. 7–12) raise against the position of those philosophers who accept Moore's verdict that F-sentences do not express propositions (Cp. also Mates 1981, pp. 36–40).

[104] In spite of this unequivocal passage, Hickman counts Bolzano among the cassantes: see his (1976, p. 400). I share the amazement voiced in Brendel (1992, p. 43).

[105] *Daß Sätze von der Form: A ist B, nie einen anderen Sinn haben, als den auch der Ausdruck: A hat b, andeutet, sofern b das zu dem Concreto B gehörige Abstractum vorstellt.*

[106] *Wenn der Buchstabe A was immer für eine Gegenstandsvorstellung, und der Buchstabe b irgendeine Beschaffenheitsvorstellung bedeutet.*

in case it 'occurs' in a proposition '*as if* it represents the object which the proposition is about (*stehet so, als ob er den Gegenstand, von welchem der Satz handelt… vorstellen sollte*)' (*WL II*, 9). [Similarly, Frege says of the Homeric name in 'Nausicaa is very beautiful' that 'it behaves as if it names a girl (*der Name … tut so, als benenne er ein Mädchen*)', and in the same vein, Quine says of all singular terms that they 'purport to refer to one and only one object' (Frege 1969, p. 133; 1979, p. 122; Quine 1974, p. 217)] So Bolzano is committed to uphold against Savonarola that (*F*) and (ψ) express propositions (see Morscher 1987).[107]

Since he endorses the principle of bivalence, he must regard those propositions as true or as false. Either option is bound to lead into well-known trouble unless that trouble is due to a misinterpretation of the pertinent sentences. Bolzano maintains that the proposition that is expressed by an utterance of (*F*) made by a person *x* at time *t* is logically equivalent (*gleichgeltend*) (Cp. *WL II* 54, 133, 201) with the proposition that is expressed if *x* utters at *t* the sentence '*Was ich so eben behaupte, erkläre ich für falsch, und behaupte es nicht*' (*WL I* 79, bottom; Bolzano's emphasis):

(F^+) What I am asserting right now I declare to be false and I do not assert it.

About what is expressed by (F^+) Bolzano says: '*Und das ist allerdings unwahr!* (And that is really not true!)'. If the principle of bivalence is correct, then lack of truth in a truth-candidate is falsity. So Bolzano declares all propositions expressed by utterances of (F^+) to be false. This verdict is plausible. A proposition expressed in context *C* by (F^+) could only be true if the proposition expressed in *C* by 'I do not assert what I am asserting right now' were true, but every proposition expressed by this simpler sentence is false: Nothing can fail to stand in a relation to something to which it does stand in that relation. Therefore, if what is said by (*F*) in *C* is logically equivalent with what is said by (F^+) in *C,* then the former is as false as the latter. So far, so good. But why should we endorse the antecedent? Bolzano's equivalence claim is just a stipulation.[108]

[107] He shows that this is a *Hintertürl* (Austrian German for a small backdoor, a loophole) through which the F-antinomy can steal into Bolzano's logic.

[108] As Spade/Read (§§ 3–4) report, some medieval authors have also argued that F-sentences are false (express falsehoods). Among the adherents of the falsity verdict are Jean Buridan [*Eleventh Sophisma*, in: Buridan 1977, pp. 143–145; 1982, pp. 86–91] and, ca. 50 years later, Paolo Veneto in the last of the 15 attempts to dissolve the F-antinomy that he presents. (The formal reconstruction of Paolo's argument in Bocheński (pp. 291–292) has come in for criticism in Brendel (1992, pp. 32–37) and Morscher (1989). Morscher uses some parts of Bocheński's apparatus for an elaborate reconstruction of Bolzano's implicit argument for the falsity verdict, but in the light of Morscher's reconstruction, too, Bolzano's equivalence thesis is nothing but a stipulation.) In his commentary (Augsburg 1516) on the *Summulae* of Petrus Hispanus, Luther's most indefatigable opponent Johann Eck also argued for the falsity verdict [cp. Rüstow, p. 116; Ashworth 1972, p. 45], and so did Tarski's 'Doktorvater' Stanisław Leśniewski (1913, pp. 77–82) [cp. Betti (2004)] and Eugene Mills (1998). Admittedly, Mills' argument for the falsity verdict presupposes that a sentence of the form 'That p is true' always expresses the same proposition as the embeded sentence alone. I have criticized this (Fregean) identity claim in (2003, pp. 34–52, 229–230, 450–452) and in (2010, pp. 410–423). Bolzano rejects the identity thesis. He maintains: For any

(The phenomenologist Hans Lipps has also criticized Bolzano's claim concerning (*F*) and (*F*⁺): '*Diese Gleichsetzung ist … verkehrt. Denn* "ich lüge*" meint doch wohl "was ich behaupte, ist nicht so wie behauptet*".' (Lipps 1923, p. 336).[109] This criticism is wrong-headed in several respects. Firstly, an equivalence thesis is not a *Gleichsetzung* (identification). An identification would be plainly false, since utterances of (*F*⁺) express propositions that are conceptually more complex than those that are expressed by utterances of (*F*): The concept *x declares y to be* ϕ is not a constituent of the latter. Secondly, Lipps' substitution of 'I am lying' for (*F*) is exegetically entirely baseless and conceptually confused. Thirdly, '*x* is lying' does certainly not mean that things are not as *x* asserts them to be. After all, not every false assertion is a lie.)

In his criticism of Savonarola, Bolzano makes a hermeneutically implausible conjecture that turns out to be philosophically fruitful. Savonarola 'seems to have thought', he suspects, that the propositions expressed (in one and the same context) by (*F*) and by

(*T*) What I am asserting right now is true

stand in contradictory opposition (*WL I* 80, first lines). But there is no textual evidence for ascribing this view to the Frate, for he does not say a word about *'Hoc* (⇄) *est verum',* and of course, he is also silent about Bolzano's (*T*) and the counterpart of (ψ):

(α) The proposition that is expressed by this (⇄) sentence is true.[110]

A fortiori, there is no basis in the text of the *Compendium* for Bolzano's subsequent surmise that (*T*) or (α) 'may have looked as absurd (*ungereimt*) to Savonarola as' (*F*) or (ψ). Admittedly, there are good reasons for thinking that something is 'fishy' about the former sentences, too. Suppose (α) expresses a proposition—call it '[α]'. Of course, from the assumption that [α] is true it does not follow that [α] is false, and from the assumption that [α] is false it does not follow that [α] is true. So there is no antinomy here. The problem with [α] is not that we cannot *consistently* assign to it either the value True or the value False—the problem is that neither assignment would have a *fundamentum in rebus*. For (almost) any proposition *x* that is a truth or a falsehood, there must be an answer to the question *why x* is true (false) that employs neither the concept of truth nor that of falsity. (I shall motivate the awkward qualification 'almost' in the next paragraph.) You obtain for '*x* is true' the kind of answer I have in mind by prefixing 'because' to a sentence that expresses the proposition *x,* and you obtain it for '*x* is false' by prefixing 'because it is not the case that' to a sentence that expresses *x* (Cp. Künne 2003, pp. 150–157). To illustrate just

p, for any *q,* (1) if it is true that *p,* then it is true because *p,* and (2) if *q* because *p,* then [*q*] is not identical with [*p*].

[109] Presumably, Edmund Husserl recommended his student Lipps to take Bolzano's reflections into account when he wrote his paper on the F-antinomy (Cp. my 1997, pp. 347–359 on Husserl's admiration for Bolzano).

[110] In the anglophone literature 'the truth-teller' is used as a nickname for (α).

the elementary case in which *x* does not itself contain the concept of truth or that of falsity: The proposition that snow is white is true *because* snow is white, and the proposition that snow is green is false *because it is not the case that* snow is green. If *x* is itself an ascription of truth or falsity then the first answer to the *why*-question is not yet the final answer, but in (almost) every case there must be an answer in which the concepts of truth and falsity no longer occur. In the case of [α] this requirement of *groundedness* is not met, and that is a defect that [ψ] shares with [α].[111] So there is a reason for regarding both as equally *ungereimt,* and this reason does not demand that we claim 'that both combinations of words [sc. (α) and (ψ)] are not real sentences (*daß beide Wortverbindungen keine eigentlichen Sätze wären*)'.

The 'almost' rider that accompanies the above explication of the notion of groundedness seems to be indispensable. Why are the propositions expressed by 'What somebody said is true (false) if, and only if, things are (not) as he or she said they are' true? One cannot answer these questions in the style characterized above without employing the concept of truth (falsity). The reason is that these propositions are truths *about* the very concepts of truth and falsity. So it is not too surprising that they are exceptions to the rule.

As I said, there is no reason to ascribe to Savonarola *any* view concerning '*Hoc* (⇄) *est verum*', (*T*) and (*a*) on the basis of the *Compendium*. Nevertheless, Bolzano's objection to the view that the propositions expressed by (ψ) and (α) form a contradictory pair and the examples he uses in support of his objection deserve close scrutiny, no matter whether they can be turned against the Frate. Bolzano presents non-paradoxical self-referential sentences as examples, and such sentences will be the main topic of the final section of this chapter.

24.3 Non-Pathological Self-Reference: Some of Its Pitfalls

24.3.1 Paradox Without Self-Reference, Self-Reference Without Paradox

Self-referentiality is not a *necessary* condition for engendering antinomies. The Parisian philosopher Jean Buridan (*fl.* 1340) has shown in his Ninth *Sophisma* that an F-antinomy can also arise in cases of *reciprocal reference*. Suppose that at the very same time at which Plato said on the Acropolis 'What Socrates is saying right now is true' and nothing else, Socrates said on the Agora nothing but this: 'What Plato is saying right now is false.' Now the question is 'whether Plato's utterance (*propositio*) is true or false, and the same question can be posed with respect to Socrates'

[111] What Mackie (1973, p. 241) says about [α] is misleading: 'We cannot decide whether the truth-teller's remark is true or false' (my italics). The defect of [α] (as well as of [ψ]) is not epistemological. For the same reason, Kripke's use of the phrase '*ascertaining* the truth-value' in his intuitive characterization of the concept *groundedness* (that I have just tried to explain in my own way) can lead astray (1975, pp. 693–694/repr. 57; my italics again).

utterance' (Buridan 1977, p. 140; 1982, pp. 78–80). The answers are thoroughly unappealing. Let us just consider the first question. What Plato said is true if, and only if, what Socrates said is true; what Socrates said is true iff what Plato said is false; hence (by hypothetical syllogism) what Plato said is true just in case it is false. Russell recounts how he was confronted for the first time with this kind of F-antinomy:

> G. G. Berry … was a man of very considerable ability in mathematical logic. He was employed in a rather humble capacity in The Bodleian [Library], his subject being one which the University of Oxford ignored. The first time he came to see me … he was bearing, as if it were a visiting card, a piece of paper on which I perceived the words: 'The statement on the other side of this paper is [true].' I turned it over & found the words: 'the statement on the other side of this paper is false'. We then proceeded to polite conversation.[112]

I shall also leave the minefield of antinomies[113] for the rest of this chapter and proceed to a set of more easily tractable problems.

Self-referentiality is not a *sufficient* condition for engendering antinomies either. One might very well think that it is agreed on all sides that many self-referential sentences are logically harmless. But in the Middle Ages there were a few *restringentes,* as they were called: For fear of paradox, they pleaded for *restrictio,* that is, global condemnation of self-reference (Spade/Read, § 2.4). The early Wittgenstein joined them. In the *Tractatus* he maintained:

> No sentence can say something about itself, *because* the sentential sign cannot be contained in itself (that is the whole of the 'theory of types'). *Kein Satz kann etwas über sich selbst aussagen, weil das Satzzeichen nicht in sich selbst enthalten sein kann (das ist die ganze 'theory of types').* (Wittgenstein 1922, 3.332; my emphasis)[114]

Is not this *ex cathedra* declaration plainly wrong? A sentence such as:

(E1) This (⇄) is an English sentence

does say something about itself, and what it says is as true as can be. In order to block the F-antinomy and its ilk, Wittgenstein's verdict imposes kin liability on thousands of innocents. To be sure, a sentential sign cannot be contained in itself, for this is just a special case of a general mereological truth: $\forall x\ \forall y$ (x is a proper

[112] Quoted after Garciadiego (1992, p. 166. Cp. Russell 1967, p. 147; Grattan-Guinness 1977, p. 50; and Philipp Jourdain's amusing presentation of the same paradox in Grattan-Guinness pp. 176–178). (Elsewhere Russell pays homage to George Godfrey Berry (1867–1928) because of what is now known as 'Berry's Paradox': see his (1908, p. 60) or Sainsbury (2009, p. 165)). As shown in Richard G. Heck (2007, p. 3), one can contract the sentences on Berry's 'visiting card' into one biconditional, namely 'The left-hand side of this (⇄) biconditional expresses a truth iff its right-hand side expresses a falsehood.' Here, reciprocal reference is turned into self-reference—nowadays Kripke's Watergate example is the best-known variant of Buridan's Ninth Sophisma.

[113] Beall and Glanzberg (2011, § 1) gives a bird's-eye view of all members of the family of antinomies whose most prominent member is the F-antinomy.

[114] The theory referred to between brackets is explained in Copi (1971). Presumably, W. regards the principle clause of 3.332 as an attempt to say something that cannot be said, as 'unsinnig' in the Tractarian sense. This move does perhaps protect it against the objection that it is self-refuting: After all, it seems to ascribe something to each and every sentence, hence also to itself.

part of $y \rightarrow x \neq y$). But this truism lends no support to the global condemnation of self-reference. As Bolzano points out, a sentence need not be a proper part of itself in order to say something about itself—containing a *singular term* that denotes it is entirely sufficient for this.[115]

24.3.2 Negation(s) of Self-Referential Sentences

Bolzano shows that *all* 'sentences whose subject or predicate contains a reference either to the whole sentence or only to some of its parts (*Sätze*, *in deren Subjecte oder Prädicate eine Beziehung auf sie selbst, oder nur auf irgend einen ihrer Bestandtheile vorkommt*)' (*WL I*, 80) have logico-semantically remarkable properties. (Admittedly, Bolzano often uses '*Sätze*' as short for '*Sätze an sich*' but at this point he has 'combinations of words (*Wortverbindungen*)', that is, sentences in mind, hence the pertinent subjects and predicates are parts of sentences.) He distinguishes here four kinds of self-reference: [1] The subject term of sentence *S* carries the self-reference, and [1a] this term denotes *S* or [1b] a part of *S,* or [2] the predicate term of *S* (i.e. the expression that follows the copula in *S*) carries the self-reference, and [2a] this term applies to *S* or [2b] to a part of *S*. The next four sentences exemplify these varieties of self-reference if (as suggested by the arrow symbol) the demonstrative pronoun or description serves the purpose of denoting the sentence in which they occur:

[1a] *This* (⇄) is a short sentence
[1b] The last word in this (⇄) sentence is a noun
[2a] The sentence in line (2a) on p. 400 is *identical with this* (⇄) *sentence*

[2b] The first letter of the alphabet is *a component of this* (⇄) *sentence*.

(The symbol '(⇄)' is not a part of the sentence—it is only meant to remind the reader of the intended interpretation.)

Now Bolzano points out: If a predication about a single object ('a is *F*', 'a is a *G*', 'a Φ−s') is self-referential, one cannot frame the *negation* of the proposition in the usual manner ('a is not *F*', 'a is no *G*', 'a does not Φ'). Let us consider his examples, or rather English variants thereof. They all have the structure [1b]. The propositions that are expressed by the next two sentences do not stand in contradictory opposition:[116]

(*S*1) The antepenultimate word in this (⇄) sentence **is** a verb
(*S*2) The antepenultimate word in this (⇄) sentence is **not** a verb.

[115] See above p. 392. More recent critics of TLP 3.332 include G.E. Moore (1962, p. 313), Martin (1967, p. 279), Mates (1981, p. 22).

[116] All references to WL that follow in this subsection are to *WL I,* 80.

The propositions expressed by (*S*1) and (*S*2) are about different words, and they are both true. So they do not contradict each other, any more than the propositions [Socrates was an Athenian] and [Kant was not an Athenian] contradict each other. 'The concept' that is expressed by the subject term 'is no longer the same (*der Begriff wird ein anderer*)', Bolzano claims, when one moves from (*S*1) to (*S*2). So by 'concept', he does not mean the linguistic meaning of the words that express it, for that remains constant during this transition. Since the verbose subject terms in (*S*1) and (*S*2) do not denote the same object, that is, the same type word,[117] they express different concepts.

Let us turn to Bolzano's second example:

(*S*3) The number of words that make up this (⇄) sentence is twelve.
(*S*4) The number of words that make up this (⇄) sentence is not twelve.

(Here it is obvious that 'Satz' in the German text is not short for 'Satz an sich', since propositions do not consist words.) If you make a count, you will realize that the propositions expressed by (*S*3) and (*S*4) are both false. They no more contradict each other than do the propositions (Diogenes was Plato's teacher) und (Socrates was not Plato's teacher).

In Bolzano's cunning examples, what *looks* like the negation of proposition *P* has the *same* truth-value as *P,* hence the appearance has got to be deceptive.[118] But of course, a similar situation can also arise when the truth-values differ. The falsehood that is expressed by:

—: This (⇄) is not an English sentence

is not the negation of the truth expressed by

(E1) This (⇄) is an English sentence,

for these propositions are about different type sentences.

In all these cases, one cannot deny in the usual manner what is said by the self-referential sentence.

And how can one deny it? This is a question Bolzano does not raise. Let me try to answer it on his behalf. There are two strategies available. The first consists in denying a proposition that is expressed by a self-referential sentence by using a

[117] Recall that we agreed earlier on to understand the demonstrative description 'this (⇄) sentence' as denoting a type sentence.

[118] I disagree with Berg's interpretation of the argument in lines 6–30 of p. 80: '[Bolzano] proves that the law of contradiction—stating that a proposition and its denial are not both false—is not applicable to self-referring propositions' [(1962, p. 60, repeated in 1985, p. 16)]. What Bolzano emphasizes is rather that some pairs of sentences which seem (line 21) to express incompatible propositions do not really do so. The use of the term 'negation' in Berg (1985, p. 16) is misleading in the same respect.

sentence that is not self-referential. Consider again Bolzano's first example. The propositions expressed by:

(*S*1) The antepenultimate word in this (⇄) sentence is a verb

and its *i*nternal *n*egation

IN(S1) The antepenultimate word in this (⇑) sentence is not a verb

do stand in contradictory opposition if the demonstrative description in *IN*(S1) is used—as the new arrow symbol is meant to suggest—to refer to a *different* sentence, namely to (*S*1). The internal negation of what is said by a self-referential sentence can only be expressed by an '*alio*-referential' sentence, because only by abandoning *self*-reference identity of *reference* can be preserved. (I shall explain in a moment the point of the adjective in the phrase 'internal negation'.)

If a sentence is self-referential, then *every* change of the wording affects the propositional content. This holds even if the wording is modified only by substitution of synonyms. The expressions 'three' and '3' have the same linguistic sense, and yet:

(*S*5) This (⇄) sentence contains three monosyllabic words.

expresses a truth, whereas

(*S*6) This (⇄) sentence contains 3 monosyllabic words.

does not. Actually, this is not more surprising than the fact that if different books are referred to then 'This book has three chapters' may express a truth while 'This book has 3 chapters' does not.

The *second* strategy for denying what is said by a self-referential sentence is less natural. It uses the concept of external negation, and it employs a stipulation. Bolzano himself carefully distinguishes the concept of external negation from that of internal negation used in the Savonarola appendix. If *P* is the proposition that Socrates is dull, then the *internal* negation of *P* is the proposition [Socrates has the property of not being dull], and the *external* negation of *P* is the proposition [*P* has the property of being false] (*WL II* 16, p. 44 ff., 63, 269, 419). We can formulate the external negation of the proposition expressed by a sentence S by putting 'It is not the case that' in front of *S*. Furthermore, let us agree on using a hook for limiting the scope of the self-referentially employed demonstrative description to the sentence that follows the hook. Now we can deny what is said by (*S*1) by using its *e*xternal *n*egation and inserting the hook:

EN(S1) It is not the case that ⌜ the antepenultimate word in this (⇄) sentence is a verb.

The insertion secures that the demonstrative description refers here to the same object, i.e. the same type sentence, as in (*S*1). But admittedly, this is just a stipulation (borrowed from Barwise and Etchemendy).[119] Its technical advantage is clear: It frees the practice of negation from the context dependency that is a characteristic of the first strategy. But I doubt that the sentence *EN*(*S*1) without hook suffers from a syntactical ambiguity that can be removed by inserting the scope delimiter.

24.3.3 Bolzano Versus Savonarola (II)

Towards the end of his reflections on Savonarola's treatment of the F-antinomy, Bolzano tries to apply his observations on negating self-referential sentences to this issue. Both are agreed that a truth-candidate is true iff it is not false. But only Bolzano thinks that utterances of (*F*) express propositions (false propositions, that is). Now he asks whether the propositions a speaker *x* at time *t* might express by uttering:

(*F*) What I am asserting right now is false
(*T*) What I am asserting right now is true

form a pair of contradictories. If the complex subject term denotes anything at all then it denotes the proposition that is expressed by the whole sentence as uttered by *x* at *t*. Now the proposition that is the topic of the (*F*)-proposition is different from the proposition that the (*T*)-proposition is about, for only the (*F*)-proposition contains the sense of 'false'. So the equiform singular terms in utterances of (*T*) and (*F*) denote different objects. But two propositions that are expressed by singular predications containing the same singular term cannot stand in contradictory opposition unless the singular term denotes the *same* object. (The proposition that Cambridge is beautiful does not contradict the proposition that Cambridge is ugly if one is about a town in England and the other about a town in Massachusetts.) Our sentence pair (*F*) and (*T*) does not satisfy the condition of co-referentiality. The same argument applies mutatis mutandis to the pair:

(ψ) The proposition that is expressed by this ($\rightleftarrows$) sentence is false
(α) The proposition that is expressed by this ($\rightleftarrows$) sentence is true.

All this is thoroughly convincing. The question how one can deny the proposition expressed by (ψ), or the proposition expressed by (α), can be answered along the same lines as in the case of (*S*1): It can be denied either internally by using an *alio-*

[119] In being a scope delimiter my '⌈' has the meaning as the arrow '↓' in the formal language constructed by Jon Barwise and John Etchemendy (1987, p. 33)—I just wanted to avoid a third kind of arrow.

referential sentence or externally by employing the scope delimiter. In order to deny the proposition expressed by (*ψ*) we cannot use (*a*),[120] but we can employ either:

IN(ψ) The proposition that is expressed by this (⇑) sentence is not false,

if the arrow points to (*ψ*),[121] or (cp. Barwise & Etchemendy 1987, p. 33; Brendel 1992, p. 136)

EN(ψ) It is not the case that ⌜ The proposition that is expressed by this (⇄) sentence is false.

Similarly, in order to deny the proposition expressed by (*α*) we can use:

IN(*α*) The proposition that is expressed by this (⇑) sentence is not true,

if (*a*) is pointed to,[122] or

EN(*α*) It is not the case that ⌜ The proposition that is expressed bythis (⇄) sentence is true.

Unfortunately, the final move in Bolzano's discussion of Savonarola is thoroughly unconvincing, and oddly enough, he himself has shown in his earlier discussion why it is unconvincing. He claims that the following two sentences do express (in the same context) a pair of propositions that stand in contradictory opposition:

(*) I do not assert what I am asserting right now
(†) I do assert what I am asserting right now.

But everything he said about the pairs (*S*1)–(*S*2) and (*F*)–(*T*) also holds of this pair. (Jan Berg put his finger on this a long time ago. Berg 1962, pp. 60–61; and 1985, p. 16.) The proposition of which speaker *x* at time t by uttering (*) says that he does not assert it clearly differs from the proposition of which *x* at t by uttering (†) would say that he asserts it: The former contains the sense of the negation operator, the latter does not. So these two propositions do not stand in contradictory opposition.

[120] As Elke Brendel (1992, p. 44) may have been the first to observe, 'Bolzano hat…den wichtigen Unterschied zwischen [dem] 'Wahrsager' und [der] Negation des 'Lügners' erkannt und damit auf eine zentrale Struktureigenschaft selbstbezüglicher Sätze hingewiesen'. (I took the liberty of inserting two pairs of scare-quotes. About the '*Lügner*' I have said enough. Her translation of the logicians' term 'truth-teller' also needs a pinch of salt (op. cit. 5), since the persons who are called *Wahrsager* in German are not called truth-tellers in English but rather soothsayers, diviners or fortune-tellers).

[121] If you want to deny what you asserted a moment ago by uttering (*F*), you can say: 'What I asserted a moment ago is not false.'

[122] If you want to deny what you asserted a moment ago by uttering (*T*), you can say: 'What I asserted a moment ago is not true.'

As the next and final sentence in Bolzano's debate with Savonarola betrays, he is untypically impatient at this point: 'But enough of such hair-splitting! (*Doch schon genug von dieser Spitzfindigkeit!*)' or to quote the March Hare, 'Suppose we change the subject. I'm getting tired of this.'

24.3.4 Deductions with Self-Referential Sentences

When harmlessly self-referential sentences occur in deductive arguments, situations arise that are at least prima facie surprising. Consider the following argument:

(A1) The first letter of the **a**lph**a**bet occurs only twice in this (⇄) sentence. So (?), it is not the c**a**se th**a**t it is not the c**a**se th**a**t the first letter of the **a**lph**a**bet occurs only twice in this (⇄) sentence.

Here we move from a truth to a falsehood, since the conclusion contains four additional occurrences of that letter. So we should better not represent this invalid argument as a substitution instance of the universally valid schema '*P, ergo* it is not the case that it is not the case that *P*'. It takes more to instantiate this schema than to contain two occurrences of the same univocal sentence.

If we were to insert the scope delimiter '⌈' between the double negation operator and the sentence we would ensure that only the embedded sentence is referred to by the demonstrative definite description, and then the conclusion of (*A*1) would be true, and the argument would be valid. But as I said before, I do not think that this manoeuvre articulates a way of understanding the original sentence: There is no scope ambiguity in the conclusion of (*A*1).

However, if a self-referential sentence is part of a *paratactic* compound of sentences, a scope ambiguity that could be removed by inserting the hook does seem to occur. Consider the following argument:

—: This (⇄) sentence contains five words. So neutrinos have mass or this (⇄) sentence contains five words.

Bill Hart maintains that its premiss is true while its conclusion is false because it consists of nine words (Hart 1970, p. 525; my arrows). But is the conclusion not syntactically ambiguous? Under one reading, the whole disjunction falls into the scope of the demonstrative description, and the argument is indeed invalid. But under another reading, only the second disjunct is demonstrated, and then the argument is valid, being an instance of the universally valid schema '*P*, *ergo Q* or *P*'. In paratactic constructions, there really is room for a narrow-scope reading of an embedded self-referential sentence.[123] The examples that I shall present all resemble (*A*1) in *not* containing such embeddings.

[123] I agree here with Barwise and Etchemendy (1987, p. 32).

If a self-referential sentence *S* expresses a truth then it is possible that the result of replacing a singular term in *S* by a co-referential term expresses a falsehood even though *S* contains no opaque construction such as quotation marks or 'believes that'. Thus, in the argument:

(*A*2) The first letter of the **a**lph**a**bet occurs twice in this (⇄) sentence. The first letter of the alphabet is the letter A. So (?), the letter A occurs twice in this (⇄) sentence.

we move from true premisses to a false conclusion (Cp. Hart (1970) p. 525/526). So we should better not regard this fallacious argument as a substitution instance of the logically valid schema 'Fa, $a=b$, *ergo* Fb'. As far as predicate letters are concerned, more is needed for instantiating this schema than: containing two occurrences of the same univocal predicate.

Here is my third example:

(*A*3) Not every word in this (⇄) sentence contains **a** consonant. So (?), at least one word in this (⇄) sentence contains no consonant.

While the premiss is true, the conclusion is false, since it contains no syllable that lacks a consonant. Hence, (*A*3) only *seems* to be a substitution instance of the logically valid schema '$\neg\forall x\,(Fx \rightarrow Gx)$, *ergo:* $\exists x\,(Fx \,\&\, \neg Gx)$'. For my final (German) example, I took my cue from Jean Buridan:[124]

(*A*4) Alle Wörter dieses (⇄) Satzes haben mehrere Silben. Also (?) kein Wort dieses (⇄) Satzes hat nicht mehrere Silben.

The premiss is true, but the conclusion is false, for it contains several monosyllabic words. So this argument cannot really be what it seems to be at first sight—a substitution instance of the universally valid schema '$\forall x\,(Fx \rightarrow Gx)$, *ergo* $\neg\,\exists x\,(Fx \,\&\, \neg Gx)$'.

In each of these cases, deductive correctness can be insured if we make the conclusion *alio*-referential by taking the demonstrative description to refer to an earlier sentence. For example, the following argument *is* deductively correct:

(*A*3*) Not every word in this (⇄) sentence contains a consonant. So, at least one syllable in this (⇑) sentence contains no consonant.

In using this strategy for securing deductive correctness, we apply the following principle: If there are several occurrences of the same demonstrative in an argu-

[124] The First *Sophisma* in Buridan (1977) p. 124; (1982) p. 42 'Omnis syllaba est plures litterae, ergo nulla syllaba est unica littera'. A literal translation of the (self-referentially understood) original would have spoilt the point. Similarly, in an English translation of (A4) that renders '*Alle Wörter dieses*…' by 'All words of this…' the premiss is no longer true, since it begins with four monosyllabic words. This translation issue is my topic in the final subsection.

ment, the argument is not deductively correct unless this expression has the same reference in each of its occurrences. In other words, in Quine's words:

> Words of ambiguous reference such as 'I', 'you', 'here', 'Smith' and 'Elm Street' are ordinarily allowable in logical arguments without qualification; their interpretation is indifferent to the logical soundness of an argument provided merely it stays the same throughout the space of the argument. (Quine 1974, § 8.)[125]

What holds for 'You live in Cambridge, hence it's not the case that you don't live in Cambrige' (one has to keep the addressee and the town constant) also holds for arguments in which a premiss contains the demonstrative definite description 'this sentence' that recurs in the conclusion: If the argument is to be correct then the reference of this expression must stay the same throughout the space of the argument. So if we interpret it as self-referential in the premiss we must *not* give it a self-referential reading in the conclusion. Suppose we replace the clumsy double-negation operator in (A1) by '¬ ¬'. Then the conclusion would be true, but the argument would remain invalid, since the reference of the singular term 'this sentence' would shift.

24.3.5 Translation of Self-Referential Sentences

Two of three intuitively very plausible constraints on translation should be violated if one has to translate self-referential sentences, and good translators are used to violate those constraints.[126] The strictures I mean are:

(I) preservation of truth-value,
(II) preservation of reference and extension, and
(III) preservation of linguistic meaning.

The requirement that the translation of a declarative sentence has to leave its truth-value unimpaired is not negotiable, I think. Mark Richard does not agree. Attacking a prominent adherent of this constraint, he argues (Richard 1997, p. 207):

> [Alonzo Church] assumes that a sentence and its [correct] translation can't diverge in truth-value, but surely [sic] this is false. 'He thinks that Phil is a groundhog' and 'He thinks that Phil is a woodchuck' may diverge in truth-value, but they are both translated by the same sentence in French, which has but a single word for the woodchuck [sc. 'marmotte'].

Sometimes the 'target language' of a translation has only one word where the 'source language' has two synonymous words. What Richard says about French also holds true for various other Romance languages—and for German, too ('Murmeltier'). If he wants to use this observation for a refutation of Church's assumption, he needs three premisses. Firstly, a sentence such as (US) might express a truth:

---

[125] The phrase 'ambiguous reference' is used here in the sense of '(contextually) shifting reference'.

[126] This subsection owes much to Bill Hart (1970) and Tyler Burge (1978).

(US) John Smith believes that groundhogs hibernate, but he does not believe that woodchucks hibernate.

Secondly, the correct translation of (US) into, say, German is

(Ger) John Smith glaubt, dass Murmeltiere Winterschlaf halten, aber er glaubt nicht, dass Murmeltiere Winterschlaf halten.

And thirdly, (Ger) always expresses a falsehood (if the negation operator governs the second conjunct as a whole). From these premisses, it follows that fulfilment of (I) is not obligatory for a good translation.

In spite of Richard's trumping 'surely' I think that only the third premiss is beyond doubt. As for the first premiss, it is at least doubtful that (US), literally understood, can be true. If the terms 'groundhog' and 'woodchuck' are synonyms (I take Richard's word for it), then the latter should be substitutable by the former in the content clause of a belief ascription, but if we make this substitution in (US), then the resulting statement is as false as (Ger). To be sure, we often let reports like (US) pass as correct, but that might only show that our practice of reporting is fairly sloppy. Perhaps (US), though literally expressing a falsehood, suggests something that is true. What can be our evidence for a report such as (US) if not the linguistic behaviour of the person reported? John Smith understands the sentence 'Groundhogs hibernate', and he is ready to affirm it, but he is not willing to assent to 'Woodchucks hibernate'. So, he is under a misapprehension as to the meaning of the term 'woodchuck', and we would do well to reformulate our report if we want it to be literally true:

(US*) John Smith believes that groundhogs hibernate, but he does not believe that the animals called 'woodchucks' hibernate.

Of course, the expressions 'groundhog' and 'animal called 'woodchuck'' do not have the same linguistic meaning (even if 'groundhog' and 'woodchuck' do).

In any case, the second premiss in my reconstruction of Richard's reasoning is plainly wrong. (Ger) is not a correct translation of (US), for, as Quine once put it, 'there can be…no stronger evidence of bad translation than that it translates earnest affirmations into obvious falsehoods' (Quine 1976, p. 113).[127] So a reasonably charitable translator would rather give up in despair than render (US) by (Ger). If she does not give up she might opt for a rendering that captures the message that is indirectly conveyed by an utterance of (US):

[127] Consider the German revenge for (US): 'Hänschen glaubt, dass seine Großmutter bis Samstag bleibt, aber er glaubt nicht, dass sie bis Sonnabend bleibt'. (Pardonably, Little Hans is under the misapprehension that 'Sonnabend' means as much as 'Sonntagabend', that is, Sunday evening.) If Richard wants to use this for an argument against Church's assumption he has to maintain that this is to be translated by the glaring falsehood: 'Hänschen believes that his grandmother will stay till Saturday, but he does not believe that she will stay till Saturday.'

(Ger*) John Smith glaubt, dass Murmeltiere Winterschlaf halten, aber er glaubt nicht, dass die Tiere, die man 'woodchucks' nennt, Winterschlaf halten.

I conclude that Richard has not shown that requirement (I) is not obligatory.

Of course, meeting this obligation is not sufficient for being a correct translation: 'Three is odd' is not an acceptable translation of 'Zwei ist gerade'. This rendering is excluded by constraint (II) that in translating one has to preserve the reference of singular terms and the extension of predicates. One is inclined to agree with Langford who took this to be an indispensable requirement: 'subject matter must remain unchanged under translation' (Cooper Harold Langford, one of the pioneers of modal logic, in (Langford 1937, p. 53)).[128] Now the pair 'The Morning Star has not much H_2O' and 'Der Abendstern hat nicht viel Wasser' complies with this demand. Nevertheless, these sentences do not seem to be good translations of each other, which gets us to the final constraint.

According to (III), a translation ought to preserve the linguistic (lexico-grammatical) meaning of the original whenever that is possible. Frege imposed a similar stricture on translation (Frege 1969, p. 222; 1979, p. 206):

> We can translate a sentence into another language. The sentence in the other language is different from the original one, for its constituents are different and are put together differently; but if the translation is correct, it will express the same sense (*Sinn*)… I call this sense a thought.

This is similar to (III), but it is not the same constraint, for Fregean *Sinn*—noëmatic sense, as one might call it, with a bow to Husserl—is not to be identified with linguistic (lexico-grammatical) meaning. The sentences 'Some people own a dog' and 'Some people own a cur' have the same truth-evaluable noëmatic sense, they express the same thought (proposition). But they differ in linguistic meaning: The word 'cur' is derogatory while the word 'dog' is not, which amounts to a difference in linguistic meaning, and this difference affects the linguistic meaning of the sentences in which they occur. The following three sentences also differ in linguistic meaning, but according to Frege they have the same noëmatic sense: 'Snow is white,' 'It is true that snow white,' 'Is snow white?' (for references and discussion see Künne 2010, pp. 410–423, 423–427, 444–454). Note that Frege only claims that preservation of noëmatic sense is a necessary condition for the correctness of a translation. So he is not committed to the bizarre contention that the English interrogative sentence 'Is snow white?' is a correct translation of the German truth-ascription 'Es ist wahr, dass Schnee weiß ist'. Nor would he have to deny that 'Manche Leute besitzen einen Köter' is a better translation of 'Some people own a cur' than 'Manche Leute besitzen einen Hund'.[129]

[128] A leading member of the 'Leipzig School' of translation science endorsed this requirement when he wrote: 'Als gesichert dürfte gelten, daß in der Translation die Invarianz in bezug auf das Denotat gewahrt bleiben muß' (Otto Kade 1968, p. 209).

[129] 'The difference between translation and original', Frege maintains, should not involve more than 'the colouring and shading that poetry and rhetoric try to give to the sense (*die Färbungen und Beleuchtungen, welche Dichtkunst [und] Beredsamkeit dem Sinne zu geben suchen*)' [1892,

Do constraints (II) and (III), and Frege's variant of (III), deserve as much respect as constraint (I)? As school children, we already learnt that the result of a word-by-word translation is not always deemed to be praiseworthy. Sometimes the result is ungrammatical, as when we render 'Mihi liber est' as *To me the book is*. Sometimes the result is not idiomatically correct: We should not translate 'Good morning!' word-by-word into Italian (Spanish, French), since native speakers of those languages do not call out to each other in the morning 'Buon mattino*!*' ('¡Buena mañana!', 'Bon matin!') but rather 'Buon giorno!' ('¡Buenos días!', 'Bonjour!'). Since this example is not a declarative sentence, it is not a case for which constraint (I), the requirement of truth-value preservation, is pertinent. Nevertheless, it already shows that sometimes the counterpart of a component of the original in a correct translation has a different lexical meaning and even a different extension: no morning is a *giorno* (*días, jour*), and no *giorno,* etc. is a morning. The smallest units in translation are not words. It leaps to the eye that sometimes the smallest unit is longer than a single sentence, as soon as we have to translate self-referential sentences.

Consider again my first example of a non-paradoxical self-referential sentence:

(*E*1) This (⇄) is an English sentence.

If we translate (*E*1) word by word into German (which can be done effortlessly in this case) and insist on a self-referential reading of the translation, we obtain:

—: Dies (⇄) ist ein englischer Satz.

This is a very bad translation, for it turns a truth into a falsehood. We can comply with the requirement of truth-value preservation if we understand the demonstrative in our first attempt as referring to (*E*1):

—: Dies (⇑) ist ein englischer Satz.

But this is not good enough either, if we come across (*E*1) in a context in which the self-referential reading is obviously intended. Then we should translate (*E*1) not only *salva veritate* but also—to honour this salvation, too, with a resounding title—*salva reflexivitate*:

(*G*1) Dies (⇄) ist ein deutscher Satz.

Note that the corresponding adjectives in (*G*1) and (*E*1) are not synonymous that the subject terms do not have the same reference and that the predicates do not have the same extension. Hence in such a case, pace Frege, the optimal translation does not express the same proposition as the original (Cp. the footnote accompanying (*A*4) in Sect. 24.3.4).

---

p. 31]. But of course, a translation of a poem that preserves only its noëmatic sense would be a very poor translation.

Now in a case like (*E*1), there is at least such a thing as *the* best translation into the target language. From the practice of translating poetry, we know that often there is no such thing. Here translators try hard to do justice to as many non-semantical features of the original as possible (such as rhyme, rhythm and alliteration, level of style and colouration), but since these features cannot jointly be respected, every translation is a compromise, and often several of these nonequivalent tradeoffs are equally laudable.[130] Now a lesson to be learned from translating self-referential sentences is that this situation can also arise in the case of texts that are entirely free of the kind of features I mentioned. Suppose a philosopher contends that a sentence may literally contain its own subject matter, and he offers as evidence:

(*E*2) The first word in this (⇄) sentence ends with an E.

Here, too, a truth is transformed into a falsehood if we translate (*E*2) word by word into German and insist on the self-referential reading of the translation:

—: Das erste Wort in diesem (⇄) Satz endet mit einem E.

Once again we jump out of the frying pan into the fire if we regard the demonstrative definite description as referring to (*E*2):

—: Das erste Wort in diesem (⇑) Satz endet mit einem E.

For here we no longer preserve the evidential point that the original has in its context. Here, too, translation should preserve truth-value as well as self-referentiality, as in:

(*G*2) *Das erste Wort in diesem (⇄) Satz endet mit einem S.*

Now the phrases 'mit einem *S*' and 'with an *E*' are certainly not synonymous. The subject terms in (*E*2) and (*G*2) do not have the same reference, and the predicates do not have the same extension. Hence, for more than one reason, these sentences do not express the same proposition. But now the original might as well be translated, again *salva veritate et reflexivitate,* by:

(*G*2') *Das zweite Wort in diesem (⇄) Satz endet mit einem E.*

Here, we have rendered 'with an *E*' by a synonymous phrase, but we have disregarded the lexical meaning of the numerical adjective in the original. The German sentences (*G*2) and (*G*2') do not have the same linguistic meaning, nor do they express the same proposition. But *ceteris paribus* there is no good reason to prefer

[130] As Frege emphasized, such features 'make the translation of poetry very difficult, indeed make perfect translation almost always impossible, for it is just in features which largely determine the poetic value that languages differ most' [Frege 1918, p. 63].

one to the other as a translation of (*E*2). This kind of indeterminacy of translation is undeniable.

Unsurprisingly, you can find evidence for this indeterminacy also in translations of Bolzano's remarks about negating self-referential sentences. For example, the German sentence rendered above by (*S*3) and in Rolf George's translation of *WL* by (*S*3'),

(*S*3) The number of words that make up this (⇄) sentence is twelve
(*S*3') The number of words in the present sentence is eleven

ends with the number word 'siebenzehn'. Both translators were justifiably unworried by the fact that their number words do not denote the same number as Bolzano's. Both renderings are self-referential sentences, both express falsehoods and they still do when the number word is preceded by 'not'. Thereby both translations preserve the point Bolzano wants to drive home with his pair of wordier German sentences, and there is no good reason to prefer one of these renderings to the other.

Sometimes, excellent translations do not comply with constraint (II), 'subject matter must remain unchanged under translation', although the sentence to be translated is at first sight *not* self-referential. Let me begin with a less than perfect example from a famous Frege translation.[131] In § 54 of *Grundlagen der Arithmetik* Frege says:

(*G*3) *Der Begriff Silbe des Wortes Zahl hebt das Wort als ein Ganzes heraus.*

Fortunately, John Austin did not render this as:

—: The concept *syllable in the word number* picks out the word as a whole.

Unlike the German word 'Zahl' the English word 'number' is not monosyllabic, so this is simply false. Austin's translation of (*G*3) is:

(*E*3)^{A} The concept *syllable in the word three* picks out the word as a whole.

This sentence is not about the same concept as (*G*3), so the subject matter is not preserved. (Of course, even if we stick to number words there are other translations that are as good as Austin's.) But in this case, the translator has an alternative option that preserves truth-value *and* subject matter, namely not to translate the word Frege refers to (without yet doing what in later writings he will do, namely putting it between quotation marks). This is Michael Beaney's rendering:

(*E*3)^{B} The concept *syllable in the word 'Zahl'* picks out the word as a whole.

So let us look for a more appropriate example.

[131] Michael Dummett used this example in his (1973, p. 372).

When Alfred Tarski, in his German monograph on the semantic conception of truth wants to explain the notion of a homophonic truth-equivalence, that is, of a biconditional in which the sentence used on the right-hand side is mentioned in the truth-ascription on the left-hand side, he gives the following example—I shall quote now *verbatim* (Tarski 1935, § 1, *sub* '(3)'):

(*G*4) 'Es schneit' ist eine wahre Aussage dann und nur dann, wenn es schneit.[132]

If the translation of (*G*4) into English is to serve the same explanatory purpose, we must not translate it as:

—: 'Es schneit' is a true sentence if, and only if, it is snowing

This biconditional is as correct as (*G*4) (if we ignore the minor blemish I mentioned), but the sentence used on its right-hand side is obviously not identical with the sentence mentioned on its left-hand side: it is an *allo*phonic truth-equivalence. When we translate 'Das Wort "Lügner" enthält einen Umlaut' into English, we must not translate the material between the quotation marks, for the word 'liar' does not contain any umlaut. But in a case like (*G*4), we *ought* to translate the material between the translation marks, and that is exactly what Tarski's translator, the British philosopher of biology Joseph Henry Woodger did (Tarski 1983, § 1, *sub* '(3)'):

(*E*4) 'It is snowing' is a true sentence if, and only if, it is snowing.

This translation flouts principle (II), since the subject matter of (*E*4) is not the German sentence that is the subject matter of (*G*4).

Prima facie, I said, (*G*4) is not self-referential. *Secunda facie,* I think, (*G*4) turns out to be self-referential in an enlarged sense of this term. The strategy the translator applied as a matter of course suggests that (*G*4) in its context contains *sotto voce* (as it were) a self-referential element:

(*E*4)+ 'It is snowing' is < sc. in the language of this (⇄) sentence > a true sentence if, and only if, it is snowing.

The assumption that in sentence (*G*4) an indexical element that is relevant to its proper understanding does not appear on the surface of the utterance or inscription is not an ad hoc hypothesis. After all, we know this phenomenon from all sorts of contexts. In an utterance of 'Now it's raining cats and dogs' the reference to the *place* of the utterance that is essential to its evaluation remains implicit, and if you

[132] A blemish of this example is that the sentence which is first mentioned and then used is indexical, but this is not relevant for the point to be made. In other papers Tarski uses 'Schnee ist weiß', and Burge who also referred to (G4) as a test case pretends that Tasrki used this sentence here as well: Burge (1978, p. 145 n).

read at Heathrow airport ‘Paris—All flights cancelled’, you will assume as a matter of course that it is *from there* that you cannot fly to Hamburg.

24.4 Appendix I (Sect. 24.1.3)

24.4.1 Eubulides, ‘The Pseudómenos’, and a T-Shirt

One of the few things Diogenes Laërtius tells us about the first philosopher whom he associates with the Pseudómenos is contained in a relative clause:

> [*Εὐβουλίδης ὁ Μιλήσιος] ὃς καὶ πολλοὺς ἐν διαλεκτικῇ λόγους ἠρώτησε* (*Vitae* II.108)

In thirteen translations of his book, one can find 13 nonequivalent renderings of this clause:[133]

> E., [**1**] qui inventa, dans la dialectique, plusieurs sortes de questions syllogistiques (de Chaufepied [?], Amsterdam 1758); [**2**] el cual inventó en la dialéctica diversas formas de argumentos engañosos (José Ortiz y Sainz, Madrid 1792); [**3**] der viele Vorträge in der Gesprächsform gehalten hat (August Christian Borheck, Wien 1807; [**4**] il quale nella dialettica inventò molte maniere di argomenti sillogistici (Conte Luigi Lechi, Milano 1842); [**5**] inventeur d’un grand nombre d’arguments sophistiques (Charles Zévort, Paris 1847); [**6**] who handed down a great many arguments in dialectics“ (Charles Duke Yonge, London 1853); [**7**] the author of many dialectical arguments in an interrogatory form (Robert Drew Hicks, Cambridge 1925); [**8**] der viele dialektische Spitzfindigkeiten aufgebracht hat (Otto Apelt [1921], neu hg. v. Klaus Reich, Hamburg 1967); [**9**] il quale svolse in forma di domanda molti argomenti dialettici (Marcello Gigante, Bari 1962); [**10**] the author of many dialectical arguments in a question and answer form (Aloysius Robert Caponigri, Chicago 1969); [**11**] der in der Dialektik viele (Fang)Schlüsse vorlegte (Fritz Jürß, Stuttgart 1998); [**12**] qui formula en dialectique de nombreux raisonnements par interrogation (Marie-Odile Goulet-Cazé, Paris 1999); [**13**] quien justamente redactó muchos argumentos dialécticos (Carlos García Gual, Madrid 2007).

It is only in [13] that the second word of the relative clause (‘also’) is translated at all, and the translation is odd enough. This omission is not much of a loss, since the point of this word in the original is unclear in any case. The derogatory colouration alone makes [2], [5] and [8] the worst translations.[134] In seven renderings, Eubulides appears as the *inventor* of the seven arguments Diogenes Laërtius is about to mention. But the text does not clearly imply this: ‘*Von wem die einzelnen Trugschlüsse wirklich zuerst aufgebracht wurden, läßt sich nicht mehr zuverlässig eruieren. Daß es Eubulides war, wird* [in *Vitae*]*, strenggenommen, nicht behauptet*’ (Döring 107; cp. Cavini 1993), 102; Spade/Read § 1 and n. 2). Only six translators take up the hint given by the verb ‘ἐ*ρωτάω*’: Eubulides is said to have presented the arguments in the form of question and answer. (*Vitae* II.106 reports that Euclides of Megara and his followers were also called dialecticians ‘because they put their arguments into the form of question and answer.’) The best translations, I think,

[133] I received help from Guillaume Fréchette, Pierluigi Minari and Alvaro Vallejo.

[134] Nevertheless, [**8**] was used in Bocheński (121).

are [9],[135] and [12]. What we are told is that Eubulides 'presented many dialectical arguments in interrogatory form'. In the case of the F-antinomy, the argumentation might have run along the following lines. *A*: 'Suppose somebody says: *Hereby I am saying something that is false*. Does he say something true or something false?' *B*'s first answer: 'What he says is true.' *A* points out that this answer has a rather unpleasant consequence. *B*'s second answer: 'What he says is false'. *A* points out that this answer, too, causes serious trouble. And he keeps *B* and many generations of logicians busy by throwing up his hands, 'Where do we go from here?'

Who invented, or discovered, the F-antinomy? Diogenes Laërtius wrote his book at least three centuries after the F-antinomy began to circulate. Perhaps, he is entirely wrong about its early history, perhaps this history began only a century after Eubulides. Cavini argues that the F-antinomy is due to the Stoic Chrysippus (*fl.* 240 BC) (Cavini 1993, p. 102). If that is correct then Diogenes Laërtius is also wrong when he claims that Theophrastus (*fl.* 340 BC) wrote three books on the antinomy, and the anecdote about Philetas (*fl.* 300 BC), as commonly understood, has to be classified as an anachronistic legend. (Since I am very fond of the epitaph, I am inclined to apply *Modus Tollens*.) As a matter of fact, Cicero never ascribes the F-antinomy to Eubulides, but he explicitly ascribes it to Chrysippus (Cicero, *Acad. Pr.* II.96). (This ascription was repeated in later centuries. Around AD 400 Jerome calls a condensed version of Cicero's conditional [1] *Chrysippeum sophisma* (Hieronymus, *Epistola LXIX, Ad Oceanum,* Sect. 2; cp. Rüstow 40, p. 103). A century and a half later an argument version of Cicero's conditional [2], 'Dico me mentiri et mentior, verum igitur dico', is referred to as '*Chrysippi syllogismus*' (in a commentary on Horace: see Rüstow 102; Hülser fr. 1215).

The ancient sobriquet of the F-antinomy has the form 'def. art. sgl. masc., followed by part. sgl. masc.'. The supressed noun is not '*λόγος* (argument) (*pace* Cavini 1993, p. 88) but '*ἄνθρωπος* (man)'. This can be seen from the analogy with other non-substantival titles in the list of arguments associated with Eubulides in *Vitae* II.108:

(1) ὁ ψευδόμενος
(2) ὁ διαλανθάνων, the Overlooked (Man)
(3) 'Ηλέκτρα, the Electra
(4) [ὁ] ἐγκεκαλυμμένος, the Masked (Man)
(5) σωρ[ε]ίτης, the Heap
(6) κερατίνης, the Horn
(7) [ὁ] φαλακρός, the Bald (Man)[136]

[135] 'svolgere' means: to develop.

[136] The arguments (5) and (7) are certainly just variants of each other and are treated thus in the literature on vagueness. The nicknames (2), (3) and (4) probably also refer to one and the same problem: 'Electra knows who Orestes is. She does not know who the masked man (the overlooked man in the corner) is. But that man is Orestes. Isn't that an inconsistent triad?' For a collection and interpretation of the ancient testimonies for (5)/(7) and (6) see my (1983). On (6) cp. also my (2013, p. 81).

Occasionally, one of the participles or adjectives in (1), (2), (4) and (7) is followed by *'λόγος*[137] That is like the transition from 'the Bald Man', for example, to 'the Bald-Man Argument'—after all, the argument may be bold, but it certainly is not bald. Jacob Bernays noticed the analogy between the non-substantival nicknames of the arguments (Bernays 1853, p. 283; duly mentioned in Rüstow pp. 41–42), and in the context of our investigation, his 1853 paper also deserves to be remembered because of the following remark (loc. cit.):

> *[Die Megariker/Eristiker/Dialektiker haben] sich bemüht, alle tieferen dialectischen Probleme in veranschaulichende Exempel zu übersetzen, ein Bemühen, das die philosophische Terminologie mit Kunstausdrücken wie Sorites u.ä. vermehrt, und der Philosophie in alten und neuen Zeiten vielen unschädlichen Spott von Seiten derjenigen zugezogen hat, die alles lästig Logische mit de[m] Spitznamen des Sophistischen abzuwehren bequem finden.*

Two years later it turned out that even a historian of logic dismissed problems like the F-antinomy as '*lästig* (tedious)'. Carl Prantl, the admirably erudite author of the *Geschichte der Logik im Abendlande*, declared: 'Wissenschaftliches Interesse für die Logik bieten natürlich diese Sophismen durchaus keines dar' (Prantl 1855, p. I, 494).[138] Even some translations of the relative clause on Eubulides in *Vitae* betray this deplorable attitude, as documented by entries **[5]** and **[8]** in my list.

Quine always refers to the F-antinomy as 'the pseudomen**on**' (Quine 1953, p. 133; 1963, p. 7 f., 17, 332), and as we saw in Sect. 24.1.3, Spade and Read believe that a remark by Seneca justifies their contention that the antinomy had the 'Greek name "pseudomenon"' (Spade/Read § 1.2). They all mistake the gender of the name of the antinomy. In the meantime this confusion has gone rampant. In August 2011, one could read a *Wikipedia* article on the 'Liar Paradox (pseudomenon in Ancient Greek)', and one could order a *Pseudomenon Phenomenon Shirt* from the online retailer ThinkGeek:

> Only while supplies last! For 2400 years we've been trying to figure out this contradiction … It's known as the 'Liar Paradox,' but the Greeks called it 'pseudomenon,' a new word you can teach your friends when they check out your shirt…This shirt reads "This statement is false." in white print on a black, 100% cotton t-shirt.

Very soon it was out of stock, so quite a few friends of proud owners of the lovely shirt may have fallen victim to a misguided linguistic lesson.

[137] As in *Vitae* VII.44, pp. 196–197. (For the occurrence of '*ψευδόμενος*' following '*ὁ σοφιστικὸς λόγος*' in the Codices of Aristotle, *EN* VII.3, 1146ª22, cp. the critical comments in Rüstow 53 n. and Crivelli 2004, p. 141 n.)

[138] Note the rhetoric: '… of course … none whatsoever …'.

24.5 Appendix II (Sect. 24.1.3)

24.5.1 On an Outstanding Dissertation and Its Author

It seldom happens that a doctoral dissertation in philosophy is praised after more than seven decades as 'the standardwork' on its topic. It did happen to an Erlangen dissertation of 1908 with the title 'Der Lügner'. The eulogist was Benson Mates, professor of philosophy at the University of California at Berkeley, well known for his work on logic and on the history of logic, epistemology and metaphysics, especially on the Stoa and on Leibniz (Mates 1981, p. 163).[139] Alexander Rüstow (1885–1963), the author of what still is the standard work on the early history of the F-antinomy, studied philosophy, classical philology and mathematics—first in Göttingen, then in Munich. In Göttingen he was one of the three founding members of Leonard Nelson's *Neue Fries'sche Schule*.[140] David Hilbert had initiated in Göttingen discussions on Cesare Burali-Forti's and Bertrand Russell's set-theoretical antinomies, and both the famous paper in which Kurt Grelling and Nelson presented the 'Heterological' Paradox (Grelling and Nelson 1908) and Rüstow's doctoral dissertation originated in the context of these discussions. On Nelson's advice, Rüstow handed in his dissertation at the University of Erlangen as an external doctoral candidate. The subtitle of his dissertation, 'Theorie, Geschichte und Auflösung des Russellschen Paradoxons', was longer than that of the published version of 1910, and it was very misleading. (As regards the announced 'dissolution' of the F-antinomy and of the set-theoretical paradoxes, Heinrich Scholz' rightly remarked that it 'deserves its name only insofar as it satisfied the author' (Scholz 1937, p. 265). This verdict was preceded by his praise of the main part of the book.) It took only a few years and Rüstow's monograph was studied in the USA (Guthrie 1914; 1915).

Before the First World War, Rüstow had begun to write a *Habilitationsschrift* on Parmenides. After the war he conquered new frontiers. He first worked at the Ministry of Economic Affairs and then as economic advisor. In the 1930s he left Germany and became professor for Economic History at the University of Istanbul. After his return to Germany he obtained (as successor of Alfred Weber) the Chair for Economic and Social Science at the University of Heidelberg (1949–1956). He became famous for his incisive criticism of *Laissez-faire* Liberalism, and he is regarded as one of the fathers of of the 'Social Market System' that shaped the economy of West Germany. As his son reports, Rüstow continued to work on his first book: a copy with extensive annotations for a new edition is part of his literary remains that are stored in the *Bundesarchiv* in Koblenz.[141]

[139] Bocheński pp. 151–152, Kneale and Kneale (1962) p. 656, Cavini (1993) p. 89 and Rescher (2001) p. 200 implicitly or explicitly concur.

[140] In English it is commonly called the Neo-Friesian School, which might evoke unwelcome associations with Friesian cattle. The school is not named after a coastal region along the North Sea but after the philosopher Jakob Friedrich Fries (1773–1843) whom I had occasion to mention before.

[141] Cp. Peckhaus (1990, pp. 131 f., 140 ff., 186–191) for an analysis of the failed solution of the semantical and set-theoretical antinomies and for further biographical details.

24.6 Appendix III (Sect. 24.1.4)

24.6.1 The Tomb of Zeus. An Ancient Controversy

When Homer and Hesiod say of the Olympian Gods that they are (live) *always,* they intend a unilateral reading of the adverb: The Olympian Gods never die, but they were born. Hesiod's *Theogony* has a lot to say about the genealogy of immortal beings. (According to *Vitae* I.111, Epimenides also wrote a theogony.) Presumably, Callimachus regarded the Cretans' claim to have the *tomb* of Zeus on their island as a manifestation of their impiety. The satirist Lucian (*fl.* AD 160) made fun of them because they are 'not ashamed of exhibiting the tomb of Zeus'.[142] The 'Church Fathers' regarded the Cretans' aspiration as a sign of disillusionment concerning the Olympian Gods. Clement of Alexandria said in his 'Exhortation to the Greeks' (Clemens 1995, ch. 2: 37, pp. 3–4; 38, 1):

> Are you looking for your Zeus? Search thoroughly not in heaven but on earth. The people of Crete will provide you with information, for there he is buried—Callimachus, in his hymns, says: 'a tomb for you, O Lord,/the Cretans built!' For Zeus is dead… even the superstitious seem … to have come to see that their conception of the Gods is erroneous.

In his 'Plea for the Christians', written ca. 177 and addressed to Emperor Marcus Aurelius and his son, Athenagoras of Athens said about lines [8]–[9] of the *Hymn to Zeus*: 'You believe, Callimachus, in the birth of Zeus, but not in his grave … for you do not know that the unbegotten God alone is eternal' (Athenagoras 1857, cap. 30: pp. 18–19, 21–22; Cp. Kokolakis 1995, p. 129). Around 248, Origen develops this into an explicit argument against the assumption of 'unilateral' eternity in the theogonies:

> On Earth, birth is the beginning of death. Now the poet [sc. Callimachus] says [of Zeus]: 'In Parrhasia it was that Rheia gave birth to you.'[143] … The poet should have realized that to the birth of Zeus in Arcadia his death corresponds as a necessary result. But Callimachus says: …

He then goes on to quote verses [6]–[8] of the *Hymn* (Origenes, *Contra Celsum,* III. 43).[144]

As for Zeus', birthplace, the Cretans seem to have preserved their old conviction: 'Mount Juktas to the south of Cnossos [is] known even to this day as the μνῆμα Διός, the burial place of Zeus; …there is an ancient sanctuary on the summit, and on the south slope a cave which has revealed evidences of an ancient worship' (Morrow 1960, p. 26).

---

[142] In his dialogue 'The Lover of Falsehoods, or the Disbeliever (*Φιλοψευδὴς ἢ Ἀπιστῶν*)', in Lucian (1913, p. 324), and at several other places.

[143] *Hymn to Zeus* [10]. Parrhasia lies in the south of Arcadia on the Peleponnese.

[144] Origenes, Contra Celsum, III. 43. On his life and work see von Campenhausen (1955, pp. 43–60).

24.7 Appendix IV (Sects. 24.1.4, 24.2.1)

24.7.1 Did the Psalmist Miss a Paradox?

The paradox hunters of the twentieth century let an opportunity slip. If Ps.-Paul missed a paradox (as Russell and Quine maintained), then so did the psalmist—and Paul the Apostle who repeated his words. If the paradox 'arises in a New Testament context' (as Saul Kripke claimed) then it also arises in the context of the Hebrew Bible. As you may have expected, I propose to apply *Modus Tollens*.

Here are a few verses from the thanksgiving Psalm 116 that contains the dictum I have in mind:[145]

[3]	The cords of death entangled me, and the fear of the netherworld caught hold of me: I was overcome by distress and sorrow.
[4]	Then I called upon the name of Yahveh, 'I beseech you, O Yahveh, save my life.' […]
[8]	You saved my life from death, my eyes from tears, and my feet from stumbling.
[9]	Now I shall walk before Yahveh in the lands of the living.
[10]	I trusted you even when I said, 'I am devastated.'
[11]	In my despair I said, 'Every man is a liar (כָּל-הָאָדָם כֹּזֵב).'
[12]	How can I ever repay to Yahveh all the good he did to me?

In the *Septuaginta,* the first Greek translation of the Hebrew Bible,[146] the exclamation in [11] is rendered by '*Πᾶς ἄνθρωπος ψεύστης*'. Paul echoes it in his *Letter to the Romans* when he says,

[Rom 3: 4]	God is truthful, but every man is a liar (*ὁ θεὸς ἀληθής, πᾶς δὲ ἄνθρωπος ψεύστης*).[147]

If you replace 'man' by 'Cretan', you obtain (an equivalent of) Epimenides' dictum.

In his 'Homilies On the Psalms' Jerome commented on *Ps* 116: 11, and of course, its echo in *Rom.* 3: 4 did not escape him (Hieronymus, *Tractatus de Psalmo CXV*, lines 62–63).[148] He begins his comments with the claim that the optimal translation of the predicate in the Hebrew original is '*mendacium* (lie)'—rather than '*mendax* (liar)', as the *Septuaginta* rendering suggests. Since this seems to be an error on his

[145] My translation follows to a large extent the *Zürich Bible* that adopts (like all Protestant and Anglican translations) the numbering of the psalms in the Hebrew Bible.

[146] On its complicated history see von Campenhausen (1955, p. 51).

[147] Since in *II Cor* 4: 13 Paul quite explicitly quotes the (incorrect) *Septuagina* translation of [10a], it is not unlikely that here he has [11b] in mind.

[148] Why 'on Ps 115'? The *Septuaginta* splits Ps 116 into two: the above verses are in Ps 114: 3–4, 8–9 and Ps 115: 1–3. The Greek numbering of the psalms was adopted in the *Vulgata* that is largely Jerome's work. He mentions the divergence in numbering: op. cit. lines 1–3. In most, cases he translated directly from the Hebrew text, but the translation of the *Book of Psalms* as contained in the *Vulgata* (with which he began his work) is only the result of correcting an older Latin rendering in the light of (Origenes' version of) the *Septuaginta*. See von Campenhausen (1960) p. 132.

part, I skip his attempt to explain [11b] under the 'lie' reading. He goes on to comment on the text under the 'liar' reading that coincides with his translation of Paul's echo of [11b] as 'omnis homo mendax'.

Before presenting his own interpretation, Jerome sketches two 'arguments of the philosophers (*syllogismi philosophorum*)' (op. cit. lines 42–43) to which such sentences give rise. The first one runs as follows:

> If what David says, 'Every man is a liar', is true then he himself is also lying (*mentitur*), since he is a man. But if he, too, is lying, then what he said, 'Every man is a liar', is not true. (op. cit. lines 36–39)

(Jerome calls the speaker David because in the Psalter many psalms, though not no. 116, are explicitly ascribed to King David.[149]) Suppose that the above translation is correct in rendering 'mentiri' by 'to lie'. Then both conditionals should be rejected: the first because the statement 'David is a liar' does not imply 'whenever David makes a statement, David lies', and the second because 'David is lying' does not imply 'What David is saying is not true' (op. cit. lines 35–42). (We have seen that philosophers of the twentieth century repeated both mistakes.) Suppose the above translation is incorrect—'mentiri' should be rendered as 'to say something that is false'. Then the first conditional should again be rejected, since 'David is a liar' does not imply 'David is saying something that is false'. So under both interpretations, the argument is fallacious.If we stick to the second reading of the verb and replace the psalmist's exclamation by 'Whatever is asserted by a man is false', we obtain a variant of the argument that leads to what I dubbed the Cretan Paradox. But Jerome would be right if he were to brush it off as irrelevant to what the psalmist said.

Jerome goes on to present the case of a speaker who says that he is saying something false. With unmistakable irony, he reports that by considering such a case

> philosophers have found out … how one can say something true and something false in one and the same utterance (*invenerunt ... philosophi ... quomodo in eodem sermone et vera quis dicat et mentiatur*). (op. cit. lines 52–54)

Under the most charitable interpretation he is gesturing here in the direction of a variant of the argument from (I), (II) and (Bivalence) to (Truth-Value Agglomeration). At any rate, he takes the philosophical 'discovery' to confirm the severe warning in the *Letter to the Colossians*: 'See to it that nobody carries you off by philosophy and vain deceit' (Paul (?), *Col* 2: 8; quoted in op. cit. lines 50–52).

Jerome's aside on the paradox-sustaining arguments of the philosophers is rather perfunctory, but he seems to me to be right in regarding those arguments as entirely irrelevant for an adequate understanding of *Ps* 116: 11. 'Let us rather maintain', Jerome recommends, 'that David said something true when he said that every man is a liar' (op. cit. lines 54–55). Even if the proposition literally expressed by [11b] is not true, the context makes it fairly clear, I dare say, what the psalmist wants to

[149] When King Saul was 'tormented by an evil spirit', young David played the 'harp' for him (*I Sam* 16: 14–23; Rembrandt 1658). In the Historical Books of the Bible King, David is said to have composed songs (*II Sam* 22) and to have introduced Temple singing (*II Chron* 23: 18, *Neh* 12: 46).

convey by his exclamation: While you can bank on Yahveh when you badly need help, men are notoriously unreliable—all too often they let you down.[150] At least the human part of this message has a good chance of being true.

24.8 Appendix V (Sect. 24.2)

24.8.1 Philosophical Writings of a Preacher of Repentance

Savonarola's *Compendium Logicae*, **[a]**, was printed for the first time in Pescia in 1492. The front page reads *Compendiu[m] logic[a]e fratris Hieronymi Savon[a]rol[a] e de Fer[r]aria ordinis pr[a]edicatorum*. This tells us that the author was a native of Ferrara and a friar of the *Ordo Praedicatorum* (O. P., Order of Preachers), i.e. of the Dominicans. In 1497, 1 year before the friar was executed on the Piazza della Signoria, his manual was published in Florence by Bartolomeo de' Libri. The book consists of ten chapters (*libri*). In Savonarola's literary remains there was an eleventh chapter that he had not seen fit for publication (first published in Savonarola 1982, pp. 161–208):

[**a***] *Centum quaestiones logicales* (Hundred Logical Problems).

Discussion of these questions, some of them metaphysical, was meant to make it easier for the novices to move on to philosophy (Savonarola 1982, p. 208).

When the handbook was published by Melchior Lott[h]er in Leipzig in 1516, the Dominican editors lengthened its title by inserting an epitheton ornans and highlighting the scope of the book: *Compendium aureum totius logic[a]e, '*Golden Handbook of the Whole of Logic'. Bolzano quotes the manual from this edition.[151] In Venice the book was printed twice under the title *Compendium logices,* first by Aurelio Pinzi in 1534 and then by Lucantonio Giunta in 1542. In the latter edition, the manual was bound together with two other books by Savonarola:

[**b**] *Compendium totius philosophiae, tam naturalis, quam moralis* (Manual of Theoretical and Practical Philosophy)

[**c**] *Opus de divisione, ordine, ac utilitate omnium scientiarum, in poeticen apologeticum* (On the Division, Order and Utility of All Sciences, In Defence of Poetry).

(The Venetian editions of the logic manual are headed '*Compendium Logic**es***'. During the Renaissance it became fashionable to Grecize names (e.g. those of philosophical disciplines). That fashion lasted long: In 1836 Friedrich Adolf Trendelenburg entitled his introduction to Aristotelian logic *Elementa logices Aristotelicae*.) In [c]

[150] In the Revised Standard Version [11b] is translated as directly expressing this thought: 'Men are all a vain hope'.

[151] The editors of Savonarola's works overlooked this edition: (1982, pp. 377–379).

the Frate takes the cognitive achievement of the art of poetry to consist in making contingent general truths credible by describing paradigmatic instances (*exempla*).

In 1596, Savonarola's manual of logic was published again in Saxony, this time in Wittenberg by Simon Gronenberg and Andreas Hoffmann. Under the Greco-Latin heading '*Epitome Logica*', it appears again as the first part of a volume that also includes [b] and [c]: *Universae philosophiae epitome & De divisione, ordine atque usu omnium scientiarum, necnon de poetices ratione, opusculum quadripartitum.* The Slovakian editor Johannes Jessenius alias Ján Jessenský took the liberty of 'emending' and 'enlarging' Fra Girolamo's text as the reader is told on the title page. (Jessenský was a philosopher and a physician who had studied in Wittenberg, Leizig and Padua. He became very famous as a surgeon and anatomist, and in 1617 he was elected rector of the Charles University in Prague. In a grim way his death resembled that of the author he had edited: Since he had sided with the Winter King, he was executed in Prague in 1621, along with 26 Bohemian noblemen, by order of Emperor Ferdinand II. Cp. *Neue Deutsche Biographie* 10 (1974, pp. 425–426).) It was in the *Lutherstadt* Wittenberg that Savonarola's manual of logic was printed for the last time before 1982 when it was critically edited within the *Edizione nazionale delle Opere di Girolamo Savonarola* that appears in Rome.[152]

Like all his philosophical writings, his logic manual fell into complete oblivion. In 1859 the author of a seminal book on Savonarola's life and work complained: '[*i suoi scritti filosofici*] *non si troveranno citati in alcun filosofo posteriore*' (Pasquale Villari 100). Two decades earlier, Bolzano had seen to it that at least for the *Compendium Logicae* this was not entirely correct. In *Wissenschaftslehre IV* 82 he again refers to the manual: here he turns an insight of the Frate against Fichte's conception of an axiom (In ch. VIII, sect. 47 (*CL* 112)). In 1897 Ernst Commer, a cofounder of German Neo-Thomism, repeatedly referred to Savonarola's manual in his textbook on Aristotelico-Thomist logic. In 1937, Heinrich Scholz noticed that Bolzano refers twice to the *Compendium Logicae* and concluded: 'so it is presumably worth a close examination' (Scholz 1937, p. 223). No such examination seems as yet to have taken place. The two standard works on the history of logic contain hardly any information about Savonarola as a philosopher. William Kneale parenthetically remarks that the Frate, a 'famous precursor of the reformation', was not a nominalist (Kneale and Kneale 1962, p. 245).[153] Józef Maria Bocheński O.P. acknowledges the existence of Savonarola's *Compendium Logicae* in the bibliography of his book and states in a footnote that his confrère was an 'important logician' (Bocheński 553, pp. 190–191). In her survey article on discussions of the F-antinomy between 1400 and 1700, Jennifer Ashworth mentions Savonarola's contribution (Ashworth 1972, p. 35, 37). Unfortunately, she misinforms her readers by claiming that Savonarola offers no argument for his view that *F*-sentences are not sentences.

[152] The two parts of [b] can be found in Savonarola (1988) as *Compendium philosophiae naturalis* (pp. 3–302) and *Compendium philosophiae moralis* (pp. 305–477), and the booklet [c] is contained in Savonarola (1982) as *Apologeticus de ratione poeticae artis* (pp. 209–271).

[153] In the *Lutherdenkmal* in Worms the precursor hypothesis is cast in bronze.

The Dominicans in Leipzig praised the *Compendium* on the front page of their edition for its conciseness and its '*mos mathematicus*'. What do they, what does Savonarola himself in the preface to his book, mean by 'mathematical procedure'? Massimo Mugnai recently raised this issue:[154]

> As is well known, humanists and Renaissance thinkers fiercely reacted against…scholastic logic…[F]rom the first half of the fifteenth century onwards the standard of rigorous reasoning began to be identified with Euclid's *Elements*, not with the Aristotelian *Organon*… In the logic books of this period it was not uncommon to find expressed the purpose of developing old-fashioned logical matters according to the rigorous method followed by mathematicians… Savonarola … for example, introducing his treatise on logic … feels obliged to criticize the obscurity and sophistry which afflict the Aristotelian logic and to state that he aims to offer a synthesis of the entire dialectic 'according to the usual practice of the mathematicians'. If we look at Savonarola's treatise, however, we easily see that … its logical content is quite traditional and does not reflect in any sense 'the usual practice of the mathematicians'.

I think this is a misunderstanding of the Frate's point. He notes that the novices are often deterred from the 'indispensible study' of logic 'by the darkness of Aristotle's books', and when he announces that he will try to help them by proceeding '*more mathematico*', he only expresses his intention to write in a manner that is as 'brief and concise, distinct and easily comprehensible' as the manner in which mathematicians are accustomed to write (*CL* 3). He is far from accusing Aristotle of *sophistry*—his frequent appeals to the *Organon* would be subverted by such an accusation.

I read Mugnai's footnote to the passage just quoted with mixed feelings: 'As a mere curiosity one may remind [sic] that Bernard Bolzano…discusses at length Savonarola's proposed solution to the liar's paradox.' Well, what one man takes to be a mere curiosity another man deems worthy of a long discussion.

Acknowledgments I presented drafts of some sections of this chapter to audiences in Köln, Frankfurt, Heidelberg, Zürich and Granada, and I am grateful for their questions and objections. Very special thanks go to Edgar Morscher and Miguel Hoeltje for taking the trouble to go through drafts of the whole chapter and to provide me with generously detailed and pointed written commentary.

References

Alexander of Aphrodisias (1891) In: Wallies M (ed) Aristotelis Topicorum libros octo commentaria (=Commentaria in Aristotelem Graeca, vol. II.2), Walter de Gruyter & Co., Berlin

Anderson AR (1970) St. Paul's Epistle to Titus. In: Martin RL (ed) The paradox of the liar. Ridgeview Publisher & Co., New Haven, pp 1–11

Aquinas (Tommasod'Aquino) (1943) Summa Theologica (Latin-German edn) vol 20. Salzburg

Aquinas (Tommasod'Aquino) (^{8}1953) Super Epistolam S Pauli ad Titum lectura. In: Cai R (ed) Super Epistolas S Pauli Lectura, vol 2. Turin

Aristotle (1890) Ethica Nicomachea In: Bywater I (ed) Oxford (ND 1963) (EN)

Aristotle (1924/^{2}1953) Metaphysics In: Ross (ed). Oxford (Metaph)

[154] *En passant*, in an illuminating study on logic and mathematics in the seventeenth century: Mugnai (2010) p. 299.

Aristotle (1949) Categoriae et Liber de Interpretatione Minio-Paluello L (ed) Oxford (Cat.; Int.)
Aristotle (1952) Meteorologica (ed/transl by: Pritchard Lee HD). London (Meteor)
Aristotle (1958) Topica et Sophistici Elenchi In: Ross WD (ed). Oxford. (Top.; SE) translations of SE: (a) Julius Hermann v. Kirchmann, Heidelberg 1883. (b) Eugen Rolfes (1922), repr. Hamburg 1968. (c) William Adair Pickard-Cambridge (1928). In: Ross (ed) The works of Aristotle, transl. into English, vol 1. Oxford, (ND 1955) (d) Pickard-Cambridge, revised by Jonathan Barnes. In: Barnes J (ed) (1984) The collected works of Aristotle, vol 1. Oxford
Aristotle (1961) Parts of animals (ed/transl by: Peck AL). London (De part anim)
Aristotle (1964) Prior and posterior analytics (Ross, Minio-Paluello (eds)). Oxford (An Pr; An Post)
Arnauld A, Nicole P ([6]1685) La Logique ou L'Art de penser. Gallimard, Paris
Ashworth EJ (1972) The treatment of semantic paradoxes from 1400 to 1700. Notre Dame J Form Log 13:34–52
Athenaeus Naucratita (1887–1890) Dipnosophistarum libri XV (Kaibel G (Ed)) 3 vols. Leipzig
Athenagoras (1857/1987) Supplicatio Pro Christianis. In: Otto C (ed) Athenagorae Philosophi Atheniensis Opera Jena Transl (1913) Bibliothek der Kirchenväter, series I, vol 12. München
Augustinus A (395) De Mendacio. In: Migne (Latina), vol 40
Augustinus A (420) Contra Mendacium. In: Migne (Latina), vol 40
Augustinus A (422) Enchiridion. In: Migne (Latina), vol 40
Bar-Hillel Y (1947) The revival of "the Liar". In: Bar-Hillel Y (1970) Aspects of language. Jerusalem, pp 244–252
Barnes J (2007) Truth, etc. Oxford University Press, Oxford
Barwise J, Etchemendy J (1987) The liar. An essay on truth and circularity. Oxford University Press, Oxford
Bayle P (1697) Euclide (natif de Mégare et disciple de Socrate). In: Bayle P (41730=51740) Dictionnaire historique et critique, vol 2. Amsterdam, pp 414–416
Beall JC, Glanzberg M (2011) Liar paradox. In: Zalta E (ed) The Stanford encyclopedia of philosophy. http://plato.stanford.edu/archives/spr2011/entries/liar-paradox/
Bellissima F, Pagli P (1995) Consequentia Mirabilis. Florence
Berg J (1962) Bolzano's Logic. Stockholm
Berg J (1985) Einleitung des Herausgebers. In: Berg J et al (ed) Wissenschaftslehre §§ 1–45, Bolzano-Gesamtausgabe, series 1, vol 11/1. Stuttgart-Bad Cannstatt, 9–20
Berg J (ed) (1992) Wissenschaftslehre §§ 349–391, Bolzano-Gesamtausgabe, series 1, vol 13/3
Bernays J (1853) Epicharmos und der Aὐξανόμενος Λόγος. Rheinisches Museum 8:280–288
Betti A (2004) Lešniewksi's early liar, Tarski and natural language. Annals Pure Appli Logic 127:267–287
Betti A (2006) On the Bolzanization of Polish thought. In: Chrudzimski A et al (ed) Brentano and Polish philosophy. Ontos verlag, Frankfurt a M, pp 55–81
Bocheński JM (1956/[4]1978) Formale Logik, OP. Freiburg-München (Bocheński)
Bolzano B (1810/1974) Beyträge zu einer begründeteren Darstellung der Mathematik. Prag/Darmstadt
Bolzano B (1834/1994) Lehrbuch der Religionswisenschaft, vol 1. Gesamtausgabe, series 1, vol 6/1
Bolzano B (1837/[2]1981) Wissenschaftslehre. Versuch einer ausführlichen und größtentheils neuen Darstellung der Logik mit steter Rücksicht auf deren bisherige Bearbeiter, 4 vols. Gesamtausgbe, series 1, vols 6–14 (1985-2000). (WL I (II, III, IV), original pagination)
Bolzano B (1851/2012) Paradoxien des Unendlichen, Leipzig. In: Tapp C (ed) Meiner, Hamburg. (New edition)
Bolzano B (1963/[2]1978) Grundlegung der Logik. Ausgewählte Paragraphen aus der Wissenschaftslehre, Band I und II (ed: Kambartel F). Hamburg
Bolzano B (1972) Theory of science (abridged. Ed, transl: George R). Oxford
Bolzano B (2002) Bernard Bolzanos Bibliothek (Berg, Morscher (eds)) Pt 2. St. Augustin (BBibl)
Bolzano B (2009) Über die Wahrhaftigkeit (18.03.1810). Berg et al (eds) Erbauungsreden 1809/1810 Bolzano-Gesamtausgabe series 2 A 17/2:291–303

Bolzano B (forthcoming) Theory of science (ed, transl: George R, Rusnock P). Oxford University Press, Oxford
Bornkamm G (1969/[7]1993) Paulus (english ed: Minneapolis 1995). Stuttgart
Brendel E (1992) Die Wahrheit über den Lügner. Walter de Gruyter, Berlin
Burge T (1978) Self-reference and translation. In: Guenthner F, Guenthner M (eds) Meaning and translation. Duckworth, London, pp 137–153
Buridanus J (1350/1977) Sophismata (Kermit Scott T (ed)). Stuttgart-Bad Cannstatt
Buridanus J (1982) John Buridan on self-reference. Chapter Eight of Buridan's Sophismata (Hughes GE (ed)). Cambridge
Buridanus J (2004) Iohanni Buridani Summularum Tractatus nonus: De practica sophismatum (Sophismata) (Pironet F (ed)). Turnhout
Campenhausen HF von (1955/[8]1993) Griechische Kirchenväter. Stuttgart. (English ed: New York 1959)
Campenhausen HF von (1960/[7]1995) Lateinische Kirchenväter. Kohlhammer Verlag, Stuttgart
Carson TL (2006) The definition of lying. Noûs 40:284–306
Cavini W (1993) Chrysippus on speaking truly and the liar. In: Döring K, Ebert T (eds) Dialektiker und Stoiker. Franz Steiner, Stuttgart, pp 85–109
Centrone S (2012) Das Problem der apagogischen Beweise in Bolzanos Beyträgen und seiner Wissenschaftslehre. Hist Philos Logic 33:127–157
Centrone S (2012) Consequentia Mirabilis, Antiskeptizismus und Antinomien. Zeitschrift für philosophische Forschung 6:539–565
Chisholm R, Feehan T (1977) The intent to deceive. J Philos 74:143–159
Church A (1946) Review of Koyré (1946). J Symbolic Logic 11:131
Cicero MT Apriora, book II (Lucullus). In: (a) Academica (Read JS (ed)). London 1885, repr. Hildesheim 1966, (b) Academicorum reliquiae cum Lucullo (Plasberg O (ed)). Stuttgart 1922, repr. 1980, (c) Hortensius, Lucullus, Academici Libri (ed, transl: Gigon O et al). Zürich 21997.
Cicero MT (1991) De divinatione/Über die Wahrsagung (ed, transl: Schäublin, C). Zürich
Cicero MT (1994) De legibus. Paradoxa Stoicorum/Über die Gesetze. Stoische Paradoxien (ed, transl: Nickel, R). Zürich
Clemens A ([1]1906/[4]1985) Stromata (Stählin O, Früchtel L (eds)). Berlin, book I. (German ed: Bibliothek der Kirchenväter, series II, vol 17, München 1936).
Clemens A (1995) Protrepticus (Marcovich M (ed)). Leiden. (German ed: Bibliothek der Kirchenväter, series II, vol 7, München 1934).
Commer E (1897/2007) Logik. Als Lehrbuch dargestellt. Paderborn/Yarmouth, Nova Scotia
Copi[lovich] IM (1971) The theory of logical types. London
Crinito P (1504/1955) De honesta disciplina. In: Angeleri C (ed) Edizione nazionale dei classici del pensiero italiano, series II.2. Roma
Crivelli P (2004) Aristotle on truth. Cambridge
Diels H, Kranz W ([11]1964) Die Fragmente der Vorsokratiker, I. Zürich-Berlin (Diels)
Diogenes L (1999) Vitae Philosophorum (Marcovich M (ed)) vol I. Leipzig (Vitae)
Döring K (ed) (1972) Die Megariker. Kommentierte Sammlung der Testimonien. B.R. Grüner Publishing Company, Amsterdam (Döring)
Döring K (1998) Eukleides aus Megara und die Megariker. In: Flashar H (ed) Grundriss der Geschichte der Philosophie, Die Philosophie der Antike, vol 2/1. Schwabe, Basel, pp 207–237
Dummett M (1973) Frege—the philosophy of language. Harvard University Press, London
Duthil Novaes C, Read S (2008) Insolubilia and the Fallacy Secundum Quid et Simpliciter. Vivarium 46:175–191
Eldridge-Smith P (2004) The Cretan Liar Paradox. In: Dobrez L et al (eds) An ABC of lying. Australian Scholarly Publishing, Melbourne, pp 72–92
Fallis D (2010) Lying and deception. Philosophers' Imprint 10(11)
Feine P, Behm J, Kümmel WG ([13]1965) Einleitung in das Neue Testament. Berlin
Frege G (1884/1986) Die Grundlagen der Arithmetik. Hamburg. English ed: (1950) Foundations of Arithmetik (transl: Austin, JL). Oxford

Frege G (1892/[8]2008) Über Sinn und Bedeutung. In: Frege G (ed: Patzig, G) Funktion, Begriff, Bedeutung. Göttingen (orig. pagination). English ed: Frege G (1997) The Frege reader (ed: Beaney M). Oxford
Frege G (1918) Der Gedanke. In: Künne W (2010) Die Philosophische Logik Gottlob Freges. Frankfurt a M. English ed: Frege, G (1997) The Frege reader (ed: Beaney M). Oxford
Frege G (1969) Nachgelassene Schriften (eds: Gabriel G et al). Hamburg. English ed: Frege G (1979) Posthumous Writings. Oxford
Frege G (1997) The Frege reader (ed: Beaney, M). Oxford
Frege G ([8]2008) Funktion, Begriff, Bedeutung (ed: Patzig G). Göttingen
Fries JF ([1]1811/[2]1819/[3]1837) System der Logik. Heidelberg. Repr: Fries JF (1971) Sämtliche Schriften, vol 7 (eds: König G, Geldsetzer L). Aalen
Garciadiego D, Alejandro R (1992) Bertrand Russell and the origins of the set-theoretic 'paradoxes'. Birkhauser, Basel
Garin E (1959/1961) Ricerche sugli scritti filosofici di Girolamo Savonarola. In: Garin E (ed) (1961) La cultura filosofica del Rinascimento italiano. Sansoni Editore, Bari, pp 201–212
Gassendi P (1658) Logica Euclidis, seu Megarica, ch III of De Logicae Origine et Varietate. In: Gassendi P (1658) Opera Omnia, vol 1. Lyon (repr. Stuttgart-Bad Cannstatt 1964), 40–42
Gealy FD (1955) The first and second epistles to Timothy and the epistle to Titus. In: The interpreter's Bible series, vol XI. New York, 343–551
Grattan-Guinness I (1977) Dear Russell—Dear Jourdain. Columbia Univ Press, New York
Grelling K, Nelson L (1908) Bemerkungen zu den Paradoxien von Russell und Burali-Forti. Abhandlungen der Fries'schen Schule 2:301–334
Gupta A, Belnap N (1993) The revision theory of truth. Bradford, Cambridge
Guthrie ER (1914) Old solutions of a new problem. Midwest Quart 1:236–241
Guthrie ER (1915) The paradoxes of Mr. Russell. With a brief account of their history. Lancaster, PA
Hart WD (1970) On self-reference. Philos Rev 79:523–528
Heck RG (2007) Self-reference and the language of arithmetic. Philosophia Mathematica III(15):1–29
Hegel GWF (1833) Vorlesungen über die Geschichte der Philosophie, vol 2 (Michelet KL (ed)). Berlin
Hickman LA 1976 Insolubilia. In: Ritter J, Gründer K (eds) Historisches Wörterbuch der Philosophie, vol 4. Schwabe, Basel, pp 396–400
Hieronymus SE ([1]1910/1918) Epistulae, vol 1. In: Hilberg I (ed) (1996) Corpus Scriptorum Ecclesiasticorum Latinorum, vol 54, 2. Wien. German edition: (1936–1937) Bibliothek der Kirchenväter, series II, vol 16. München
Hieronymus SE (1958) Tractatus de Psalmo CXV. In: Morin G, Capelle B, Fraipont J (eds) Tractatus sive homiliae in Psalmos. In: Corpus Christianorum, Series Latina, vol 78. Turnhout. English ed: Homily 40. In: (1964) The Homilies of St. Jerome, vol 1 (=Fathers of the Church, vol 48). Washington
Hieronymus SE (2003) Commentary on 'Letter to Titus'. In: Hieronymus SE (2003) Commentarii in epistulas Pauli apostoli ad Titum et ad Philemonem. In: Federica Bucchi F (ed) Corpus Christianorum, Series Latina, vol 77 C. Turnhout, 1–73 (Comm)
Hülser K (1988) Die Fragmente zur Dialektik der Stoiker. Neue Sammlung der Texte. Mit deutscher Übersetzung und Kommentar, vol 4. Stuttgart-Bad Canstatt. (Hülser)
Kade O (1968) Kommunikationswissenschaftliche Probleme der Translation. In: Wilss W (ed) (1981) Übersetzungswissenschaft. Darmstadt, 199–218
Kallimachos (2004) Werke. Griechisch und deutsch (ed, transl: Asper M). Darmstadt
Kant I (1797) Über ein vermeintes Recht aus Menschenliebe zu lügen'. Berlinische Blätter. Blatt 10:301–314
Kneale WC (1971) Russell's paradox and some others. Br J Philos Sci 22:321–338
Kneale WC (1972) Propositions and truth in natural languages. Mind 81:225–243
Kneale M, Kneale WC (1962) The development of logic. Oxford University Press, Oxford
Kokolakis M (1995) Zeus´ Tomb - An Object of Pride and Reproach. Kernos, Revue internationale et pluridisciplinaire de religion grecque antique (Liège) 8:123–138

Koyré A (1946) The liar. Philos Phenomen Res 6:344–362
Koyré A (1947a) Épiménide le menteur. Hermann, Paris
Koyré A (1947b) Reply (to Bar-Hillel). Philos Phenomen Res 8:254–255
Kripke S (1975) Outline of a theory of truth. J Philos 72:690–716. Repr: Martin RL (ed) (1984) Recent essays on truth and the liar paradox. Oxford, 53–81. Repr: Kripke S (2011) Philosophical troubles, Oxford. 75–98
Künne W (1982) Megarische Aporien für Freges Semantik. Semiotik 4:267–290
Künne W (1997) "Die Ernte wird erscheinen…" Die Geschichte der Bolzano-Rezeption. In: Künne W (2008) Versuche über Bolzano/Essays on Bolzano. Academia Verlag, Sankt Augustin, pp 305–404.
Künne W (1999) Über Lug und Trug. In: Künne W (2008) Versuche über Bolzano/Essays on Bolzano. Academia Verlag, Sankt Augustin, pp 121–156
Künne W (2003) Conceptions of truth. Oxford University Press, Oxford
Künne W (2008) Versuche über Bolzano/Essays on Bolzano. Academia Verlag, Sankt Augustin
Künne W (2010) Die Philosophische Logik Gottlob Freges. Klostermann, Frankfurt a M
Künne W (2013) Epimenides und andere Lügner. Klostermann, Frankfurt a M
Langford CH (1937) Review of EW Beth, the significs of pasigraphic systems. J Symbolic Logic 2:53–54
Lešniewski S (1913) A critique of the logical principle of excluded middle. In: Lešniewski S (1992) Collected Works, vol 1. Dordrecht, 47–85
Liddell HG, Scott R (1940) A Greek-English Lexicon (revised and augmented throughout Jones HS, Sir). Oxford University Press. Oxford (LSJ)
Lipps H (1923) Bemerkungen zur Paradoxie des "Lügners". Kant-Studien 28:335–339
Lucian of Samosata (1913) Works, vol 3 (ed, transl: Harmon AM). London
Łukasiewicz J (1915) O nauce (On Science). In: Łukasiewicz J (1998) Logika i metafizyk (Jadacki JJ (ed)). Warsaw, 9–32
Lycan WG (2010) What, exactly, is a paradox? Analysis 70:615–622
Mackie JL (1973) Truth, probability and paradox. Oxford University Press, Oxford
MacIver AM (1939) More about some old logical puzzles. Analysis 6:63–68
Martin RL (1967) Toward a solution to the liar paradox. Philos Rev 76:279–311
Martin RL (ed) (1970) The paradox of the liar. Yale University Press, New Haven
Martin RL (ed) (1984) Recent essays on truth and the liar paradox. Oxford University Press, Oxford
Mates B (1981) Sceptical essays. University of Chicago Press, Chicago
Michael Ephesios [Ps.-Alexander] (1898) Aristotelis Sophisticos Elenchos commentarium, (=Commentaria in Aristotelem Graeca, Vol II.3) (ed Wallies M). Berlin
Migne JP (ed) (1844–1855) Patrologiae cursus completa, Series Latina. Paris
Migne JP (ed) (1857–1866) Patrologiae cursus completa, Series Graeca. Paris
Mills E (1998) A simple solution to the liar. Philos Stud 89:197–212
Moore GE (1942) An autobiography. A reply to my critics. In: Schilpp PA (ed) (1942) The philosophy of GE Moore. Evanston, 3–39:535–677
Moore GE (1948/1949) The Epimenides. In: Moore GE (1962) The commonplace book of GE Moore. 1919–1953 (ed Lewy C). George Allen & Unwin, London, 378–384
Moore GE (1962) The commonplace book of GE Moore. 1919–1953 (ed Lewy C). George Allen & Unwin, London
Morrow GR (1960) Plato's Cretan city. A historical interpretation of the laws. Princeton University Press, Princeton
Morscher E (1987) Hintertürln für Paradoxien in Bolzanos Logik. In: Morscher E (2007) Studien zur Logik Bernard Bolzanos. Academia Richarz, St. Augustin, pp 159–167
Morscher E (1989) ZuBLösungderLügner-Paradoxie. In: Morscher E (2007) Studien zur Logik Bernard Bolzanos. Academia Richarz, St. Augustin, pp 149–157
Morscher E (2007) Studien zur Logik Bernard Bolzanos. Academia Richarz, St. Augustin
Mugnai M (2010) Logic and mathematics in the seventeenth century. Hist Philos Logic 31:297–314
Nestle E, Aland K (eds) ([1]1898/[27]1993) Novum Testamentum Graece et Latine. American Bible Society, Stuttgart

Novum Testamentum Tetraglotton (1858) repr. Zürich 1981
Origenes (2001) Contra Celsum libri VIII (Marcovich M (ed)). Leiden. German ed: (1926) Bibliothek der Kirchenväter, series II, vols 52–53. München
Peckhaus V (1990) Hilbertprogramm und Kritische Philosophie. Vandenhoeck & Ruprecht, Göttingen
Peirce CS (1901) Insolubilia. In: Peirce CS (1932) Collected papers, vol 2 (Hartshorne C, Weiss P (eds)). Cambridge, pp 370–371 (§ 618)
Plato (1900–1907) Opera (Burnet J (ed)). Oxonii, Oxford
Poliziano A (1553) Epistularum libri XII. In: Poliziano A (ed) (1553) Angeli Politiani Opera Omnia, Basel. Repr: Poliziano A (1971) (Maier I (ed)), vol 1. Turin
Prantl C (1855–1870) Geschichte der Logik im Abendlande, 4 vols. München
Priest G, Berto F (2010) Dialetheism. In: Zalta E (ed) The Stanford encyclopedia of philosophy. http://plato.stanford.edu/entries/dialetheism/
Prior AN (1958) Epimenides the Cretan. In: Prior AN (1976) Papers in logic and ethics. London, pp 70–77
Prior AN (1961) On a family of paradoxes. Notre Dame J Form Logic 2:16–32
Quine WVO (1953) Notes on the theory of reference. In: Quine WVO (21961) From a logical point of view. New York, pp 130–138
Quine WVO (1962) The ways of paradox. In: Quine WVO (1976) The ways of paradox and other essays (2nd rev edn). Harvard University Press, Cambridge, pp 1–18
Quine WVO (1974) Methods of logic (3rd rev edn). London
Ramsey FP (1925) The foundations of mathematics. In: Ramsey FP (1931) The foundations of mathematics and other logical essays (Braithwaite RB (ed)). Martino Fine Books, London, pp 1–61
Read S (s. Dutilh; Spade.)
Rescher N (2001) Paradoxes. Their roots, range, and resolution. Open Court, Chicago
Richard M (1997) Propositional attitudes. In: Hale B, Wright C (eds) (1997) A companion to the philosophy of language. Wiley-Blackwell, Oxford, pp 197–226
Russell B (1905) On denoting. In: Russell B (1956) Logic and knowledge (Marsh RC (ed)). Spokesman Books, London, pp 41–56
Russell B (1906) On "Insolubilia" and their solution by symbolic logic. In: Russell B (1973) Essays in analysis (Lackey D (ed)). London, pp 190–214
Russell B (1908) Mathematical logic as based on the theory of types. In: Russell B (1956) Logic and knowledge (Marsh RC (ed)). Spokesman Books, London, pp 59–102
Russell B (1918) The philosophy of logical atomism. In: Russell B (1956) Logic and knowledge (Marsh RC (ed)). Spokesman Books, London, pp 177–281
Russell B (1940) An inquiry into meaning and truth. Spokesman Books, London
Russell B (1956) Logic and knowledge (Marsh RC (ed)). Spokesman Books, London
Russell B (1959) My philosophical development. Spokesman Books, London
Russell B (1967) The autobiography of Bertrand Russell, vol 1, 1872–1914. Routledge, London
Russell B (1977) Letter to Philip Jourdain (28.04.1905). In: Grattan-Guinness I (1977) Dear Russell—Dear Jourdain. Columbia University Press, New York, 44
Russell B, Whitehead AN (1910) Principia Mathematica, vol 1. Cambridge
Rüstow A (1910) Der Lügner. Theorie/ Geschichte und Auflösung. Leipzig. Repr. New York 1987; (ed. Marion Soreth) Köln 1994. Cp. Appendix II. [Rüstow]
Rüstow A (1911) Erklärung (zur Kritik von Paul Maas). Byzantinische Zeitschrift 20:628–663
Sainsbury RM (^{3}2009) Paradoxes. Cambridge University Press, Cambridge
Saul J (2012) Just go ahead and lie. Analysis 72:3–9
Savonarola G (1492) Compendium logicae. In: Savonarola G (1982) Opere, vol 22, Scritti filosofici, 1 (Garfagnini G, Garin E (eds)). Roma 3–208. Cp. Appendix V. [CL]
Savonarola G (1982) Opere, vol 22, Scritti filosofici, 1 (Garfagnini G, Garin E (eds)). Roma
Savonarola G (1988) Opere, vol 24: Scritti filosofici, 2 (Garfagnini G, Garin E (eds)). Roma
Schnelle U (^{6}2007) Einleitung in das Neue Testament. Göttingen

Schnieder B (2007) Mere possibilities: a Bolzanian approach to non-actual objects. J Hist Philos 45:525–550
Scholz H (1937) Die Wissenschaftslehre Bolzanos. Eine Jahrhundert-Betrachtung. In: Scholz H (1969) Mathesis Universalis (Hermes H et al (eds)). Darmstadt, pp 219–267
Seneca LA (2007) Epistulae morales ad Lucilium/Briefe an Lucilius, vol 1 (ed, transl. Fink G). Düsseldorf
Sobel JH (1992) Lies, lies, and more lies: a plea for propositions. Philos Stud 67:51–69
Sorensen R (2003) A brief history of the paradox. Oxford University Press, Oxford
Sorensen R (2007) Bald-faced Lies. Pacific Philos Quart 88:251–264
Spade PV (1973) The origins of the medieval insolubilia-literature. In: Spade PV (1988) Lies, language and logic in the late middle ages. Variorum, London, ch II., 292–309
Spade PV (1988) Lies, language and logic in the late middle ages. Variorum, London
Spade PV, Read S (2009) Insolubles. In: Zalta, E (ed) The Stanford encyclopedia of philosophy. http://plato.stanford.edu/archives/win2009/entries/insolubles/(Spade/Read)
Stock StG (1908) Stoicism. London
Tarski A (1935) Der Wahrheitsbegriff in den formalisierten Sprachen. In: Tarski A (1986) vol 2, 51–198. English ed: The concept of truth in formalized languages. In: Tarski A (21983) Logic, semantics, metamathematics. papers from 1923 to 1938 (ed, transl: Woodger JH). Indianapolis
Tarski A (1944) The semantic conception of truth and the foundations of semantics. In: Tarski A (1986) Collected Papers, vol 2 (ed: Givant S et al.). Basel/Boston/Stuttgart, pp 661–699
Tarski A (1969) Truth and proof. In: Tarski A (1986) Collected papers, vol 4 (ed: Givant S et al.). Basel/Boston/Stuttgart, pp 399–423.
Tarski A (1986) Collected papers 4 vols (ed: Givant S et al.). Basel/Boston/Stuttgart
Theodoret of Cyrus (1926) Historia Ecclesiastica. In: Migne JP (ed) (1857–1866) Patrologiae cursus completa, Series Graeca, vol 82. Paris. German (ed): Bibliothek der Kirchenväter, series I, vol 51. München
van Fraassen BC (1968) Presupposition, implication, and self-reference. J Philos 65:136–152
Villari P (1859) La storia di Girolamo Savonarola e de' suoi tempi, vol 1. Florence
Williams B (2002) Truth and truthfulness. Princeton University Press, Princeton
Wittgenstein L (1922) Tractatus Logico-Philosophicus. Frankfurt a M (1989)

Chapter 25
Constitutive Versus Normative Accounts of Speech and Mental Acts

Manuel García-Carpintero

Abstract At the end of his "Promisings and other Social Acts: Their Constituents and Structure," Kevin Mulligan briefly considers the question of the normativity of speech and mental acts. This has been a matter hotly debated in recent years; several authors (including Kathrin Glüer-Pagin and Åsa Wikforss) have contended that, properly understood, superficially looking normative notions that we deploy in characterizing such acts should be understood in a constitutive, nonnormative sense (in contrast with views such as the one recently defended by Tim Williamson about assertion, on which this is a constitutively normative act). The goal of my chapter would be to critically examine these suggestions and defend a normative account.

Keywords Norms of meaning · Speech acts · Normativity · Rule following · Meaning conventions

25.1 Introduction

In different places in his work (for instance, in the very last paragraph in Mulligan (1987), throughout Sect. 3 in Mulligan (1999), and in the talk "Against Rampant Normativism" given in Parma, 2000), Kevin Mulligan considers and suggests a view recently advocated in a more committed way by Glüer and Wikforss (2009a, pp. 48–52).[1] This is the claim that constitutive accounts of meaning (force, content, or both; mental, linguistic, or both), i.e., accounts in terms of "internal relations," are at odds with normative ones: A constitutive account fully explains any intuitive sense that "oughts" are in place when it comes to meaning, and is in fact incompatible with a further explanation in terms of genuinely prescriptive oughts. Glüer

[1] While Mulligan discusses both normativity in language and cognition, Glüer and Wikforss (2009a) restrict their discussion to the latter case (although it is pretty clear they would extend their skepticism to linguistic normativity). Here, I will discuss the broader picture, without paying attention to the otherwise important differences between the two cases.

M. García-Carpintero (✉)
LOGOS-Departament de Lògica, Història i Filosofia de la Ciència, Universitat de Barcelona, Barcelona, Spain
e-mail: m.garciacarpintero@ub.edu

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_25,

and Wikforss (2009, p. 48) put it like this: "there is a clear sense in which states or performances that are internally related *cannot* stand in normative relations"; in a less committed way, Mulligan suggests: "perhaps the time has come to consider the view that semantic and cognitive relations are internal but not normative" ("Against Rampant Normativism," § 6).

Both Mulligan and Glüer and Wikforss mention Frege as a predecessor of the view they put forward. The main reason Glüer and Wikforss (2009a, p. 49 ff) provide in favor of the view appeals to the point that "oughts not only imply cans, they also imply the possibility of *violation,* of what we could call 'forbidden combinations'"; they go on to show convincingly that, even though any plausible claim about internal relations in this field—say, *modus ponens*-shaped transitions for possessors of the conditional concept, transitions from perceptual experience to belief regarding perceptual content, etc.—should allow for the possibility of exceptions, such exceptions do not adequately count as the needed *violations*. Mulligan (1999, § 3) suggests this point too, in that he emphasizes that rule following manifests itself primarily in avoidance of rule breaking: Where rules "rule, the possibility that rule-breaking rules should not be ruled out." In addition, he appeals to a disparity between deontic notions and the cognitively constitutive properties he takes to be fundamental: "The properties of being obligatory and of being allowed are properties which do not admit of degrees…[but] prima facie justification admits of degrees" (Mulligan 1999, § 3). In this note, I will only discuss the former, more worrying point, on the assumption that the latter admits the reply that it is not just internal relations of justification that are determined by meaning-constitutive facts. I will argue that, although the point is correct as far as it goes, the friend of a fully normative account of the relevant notions should not worry.

25.2 Norms and Meaning

Famously, Wittgenstein's later philosophy establishes a constitutive connection between *meaning* (*linguistic, or mental*) and *norms* or *rules*; Kripke's (1982) widely debated interpretation has further encouraged discussion about the role of normative reasons in characterizing languages. Kripke's presentation appeals to rules or norms such as these:

> (Plus) If one means *addition* by "+," one ought (to answer "125" if asked "68+57?").
> (Circ) If one means *being circular* by "circular," then one ought not (to apply "circular" to *o* if *o* is not circular) and one is permitted (to apply "circular" to *o* if *o* is circular).

The ensuing discussions have evinced to what extent almost everything in this debate is highly controversial; here, I will limit myself to developing the line that I find compelling, without going in any depth into those debates.[2]

[2] Glüer and Wikforss' (2009b) is a good introduction that gives an accurate idea of the different controversies.

I have phrased the illustrative examples above by explicitly providing scope indications so that the obligation in (plus) mentions an act, *answering,* which I understand to involve *asserting/judging*; so does the one in (circular) under the interpretation I intend for *applying,* which I also take to be a form of assertion or judgment. This is because I agree that, in an alternative interpretation on which "to apply" just means *predicating*—cf. Glüer and Wikforss (2009b, § 2.1.1)—we do not have, here, genuine rules giving normative reasons for agents to act. Let me elaborate on this.

Most speech acts have representational contents, which can be shared by acts of different illocutionary types. Representational contents—*propositions,* in one construal of them—encode correctness conditions with respect to different possible worlds: Conditions that (putting aside necessarily true or necessarily false contents) obtain with respect to some worlds, perhaps the actual world among them, and not with respect to others. But the notion of "correctness" here at stake is an etiolated one for present purposes. I can assert *p,* or order *p,* or (in fiction-making mood) propose that you imagine *p*. Arguably, that *p* is not the case in the actual world (not just *now,* but atemporally, speaking *sub specie æternitate*) does not make the order thereby incorrect (if it was one worth giving, a good even if ultimately unsatisfied try), and certainly does not make the act of fiction-making incorrect; but it does make the assertion incorrect. Similarly, that *p* is (also atemporally) the case in the actual world does not make the order correct—the order might have been a stupid one, or one given without any proper authority, thus, providing no genuinely normative reason for the recipient to act; nor, again, does it thereby make the act of fiction-making correct: *p* might well be a totally uninteresting thing for anybody to imagine. Arguably, however, it might be enough for making the assertion correct. Hence, the correctness conditions encoded by propositions do not constitute *normative reasons* in the sense that interests us: reasons for a rational subject to act. They merely affect *a division* of representational contents into two classes (with respect to any possible world): those obtaining in, or correctly representing the world, and those not obtaining.[3] That a representational act represents an obtaining proposition does not, by itself, furnish any agent with a particular normative reason to act; for

[3] I think that when ordinary speakers find the likes of (circular) intuitively compelling, they are considering only the assertion interpretation, not the alternative predication interpretation. Note that, in that alternative understanding, we are under the relevant etiolated "obligation" of "applying" "circular" to *N* not just when we say "*N* is circular" (which is what first comes to mind but does not allow us to distinguish the two interpretations, because here we are also applying "circular" to *N* in the act sense) but also (under the scope of the relevant operators) when we say "it is not the case that *N* is circular," "*N* is circular or it is not," "Peter said that *N* is circular" or "*N* is circular, I imagine"; for in all these cases we still are (under the scope of the operators) predicating "circular" of *N*. If so, the fact that we may also find (circular) correct under the predication interpretation when we fully grasp the theoretical notion of representational content (which I grant) is I think irrelevant for purposes of philosophical theorizing, and thus I do not think that that interpretation of meaning-normativity is adequate to play the role it has in Kripkenstein's rule-following considerations.

such a reason to be determined we need, in addition, the illocutionary type—the "point" of the act.[4]

I take it that rules like (plus) and (circular) are intuitively "primitively compelling," to use Peacocke's (1987) terminology.[5] Is there an explanation for why this is so? Such an explanation would appeal, firstly, to accounts of speech and mental acts in terms of constitutive rules, along the lines of Austin's (1962), Searle's (1969), Alston's (2000), or Williamson's (1996/2000) for the specific case of assertion; secondly, to a view of meanings of natural language sentences and mental states as *act potentials,* as also developed by Alston (2000)—for the linguistic case—and others. I will now briefly outline both ideas.

Williamson (1996/2000) claims that the following norm or rule (the *knowledge rule*) is constitutive of assertion, and individuates it:

(KR) One must ((assert *p*) only if one knows *p*).

The obligation (KR) imposes is not *all things considered,* but prima facie; in any particular case, it can be overruled by stronger obligations imposed by other norms.[6] Now, in the course of the debate that Williamson's proposal has engendered, other writers have accepted the view that assertion is defined by constitutive rules but have proposed alternative norms; thus, Weiner (2005) proposes a *truth* rule, (TR), and Lackey (2007) a *reasonableness* rule, (RBR):

(TR) One must ((assert *p*) only if *p*).
(RBR) One must ((assert *p*) only if it is reasonable for one to believe *p*).

As a first motivation for his account, Williamson (1996/2000, p. 252) mentions intuitive conversational patterns: We challenge assertions politely by asking "How do you know?" or, more aggressively, "Do you know that?" (Williamson 1996/2000, p. 252). Austin (1962, p. 138) already pointed out these patterns:

> [I]t is important to notice also that statements too are liable to infelicity of this kind in other ways also parallel to contracts, promises, warnings, &c. Just as we often say, for example, 'You cannot order me', in the sense 'You have not the right to order me', which is equivalent to saying that you are not in the appropriate position to do so: so often there are things you cannot state—have no right to state—are in no position to state. You *cannot* now state

[4] This is also, I take it, Dummett's (1978) main reason why "deflationary" definitions of truth constrained only to generate all true instances of (*T*) above (be they for linguistic items, as in that schema, or directly for propositions) do not suffice to (and perhaps are then unnecessary) to characterize truth, understood not as a property of representational contents but of assertions themselves—which arguably are the intuitively primary truth-bearers, for the sort of consideration invoked in the previous footnote.

[5] More in line with the goals of that paper, in stating analogous rules for expressions that are used to build complex sentences, such as logical constants (negation, implication, quantification, predication), we would invoke fundamental *argumentative transitions* (*modus ponens* in the case of implication) instead of acts such as *answering* or *applying*. I will develop this point further below.

[6] In criticizing normative accounts of acts such as assertion, Judith Thomson (2008, Chap. VI) decisively ignores this fact; additionally, she relies on the notion of correctness ("external correctness," in her terms) for contents of representational acts that was shown before not to be properly normative.

> how many people there are in the next room; if you say 'There are fifty people in the next room', I can only regard you as guessing or conjecturing.

As Hindriks (2007) notes, these facts about our practices of appraising assertions are by themselves insufficient to justify normative accounts. For we also evaluate assertions relative to (invoking Rawls' well-known distinction) merely *regulative* norms, norms that regulate, relative to certain purposes, acts in themselves *constitutively* nonnormative—for instance, as witty, polite, or well phrased. Hindriks shows that norms for assertion could be merely regulative of a constitutively nonnormative practice, definable in the motivating reasons Gricean account that Bach and Harnish proposed, GA below ('R-intending' there is to be explicated in terms of Gricean *communicative intentions*). The regulative norms in question would then be derived from an ultimately *moral* sincerity rule such as (SR):

> (GA) To assert *p* is to utter a sentence that means *p* thereby R-intending the hearer to take the utterance as a reason to think that the speaker believes *p*.
> (SR) In situations of normal trust, one ought to be sincere.

I (2004) have argued, however, that there are considerations against nonnormative accounts such as (GA), and in favor of normative accounts stronger than those provided by (TR) or (RBR). Firstly, there are well-known objections to Gricean accounts of speaker meaning in general, of which (GA) is a special case for assertoric meaning, which strongly suggest that normative accounts are preferable (Vlach 1981; Alston 2000, Chap. 2; Green 2007, Chap. 3). Thus, the clerk in the information booth uttering "The flight will depart on time," or the victim saying to his torturer "I did not do it," or any of us uttering to our neighbor in the lift "nice weather, isn't it?," may well lack the Gricean intentions that (GA) requires for them to assert but they are asserting all right; any normative account would capture this, for, no matter their intentions, they are still committed to knowing what they say (or having justification for it, or being truthful, whatever the proper rule is). In the second place, there are situations in which we may have overwhelming prudential or moral reasons to violate (SR), i.e., not to be sincere. If the "regulative rules" account were correct, any sense that we are violating a prima facie norm—even if *all things considered* we are doing the right thing—should vanish in those cases; but it does not, or at least it does not according to intuitions many of us share—exactly as it happens in analogous cases involving promises, as Rawls (1955) pointed out in his influential argument against utilitarian "regulative rules" accounts of them.

Williamson (1996/2000) provides additional justification for his specific normative proposal: First, the account explains what is wrong in a version of Moore's paradox with "know'"instead of "believe": *A, and I do not know that A* (Williamson 1996/2000, pp. 253–254). Second, mathematics provides for formal situations where the speaker's sensitivity to the norms of assertion is highlighted; in those situations, being warranted to assert *p* appears to go hand in hand with knowing *p*. Third, an account based on TR seems at first sight preferable: given that the truth rule is satisfied whenever the knowledge rule is, but not the other way around, it provides for a practice with fewer violations of its governing rule; some evidential rule could then be explained as derived from TR, and considerations not specific to

assertion. However, the truth rule does not individuate assertion; alternative speech acts like conjecturing, reminding, or swearing also involve a truth rule (Williamson 1996/2000, pp. 244–245). Moreover, reflection on lotteries (cases in which, knowing that you hold a ticket in a very large lottery, I assert "your ticket did not win" only on the basis of the high probability of the utterance's truth) question the validity of any such alleged derivation (Williamson 1996/2000, pp. 246–252). Finally, intuitions about many cases in which we assert without knowing can be made compatible with the view. In some cases, it is reasonable for us to think that we know, even if we do not; what we do is not permissible, but it is, as we feel, exculpable. In some cases, additional values (saving someone from danger, enjoying a relaxed conversation) are at stake, allowing again for exculpation based on their contextual relative strength (Williamson 1996/2000, pp. 256–259).

Speech acts like assertion are thus normative; they are constituted by rules such as (KR). This applies both to those done by resorting to purely conventional means, if there are any, and also to those done in an indirect way, having recourse to the sort of pragmatic mechanism that Grice (1975) famously characterized as *conversational implicature,* as for instance when we "ask," "who the hell would want to see a film with that plot?" A full explanation of our intuitive feeling concerning the appropriateness of (plus) or (circular) is provided by a view of what natural languages are for such as Alston's (2000), according to which the literal, primary meanings of sentences in natural language are *speech act potentials*.

Why "potentials?" This is one of the points at which important differences between the linguistic and the cognition case come up. Consider utterances of "he is hungry." This is a declarative sentence, and it is reasonable to argue that such sentences are used by default to make assertions. If so, and according to rules such as (KR) or (TR), in uttering it with its default meaning, a speaker commits himself to knowing (or to the truth of) a certain proposition. Intuitively, to identify such a proposition we need information about the context in which the utterance is made; knowledge of English tells us that the referent of "he" should be some male made salient ("demonstrated") by the speaker; but we need further information about how such salience is established in the context, in order to identify it. Recent debates in linguistics and the philosophy of language about the semantics/pragmatics distinction suggests that this point is widespread in natural languages. Literal and direct utterances of "it is raining" in a context commit their utterers to propositions concerning specific times and places; of "no one showed up at the party," to propositions concerning specific domains of discourse; of "Peter is tall," to propositions concerning specific tallness standards.

Kaplan (1989) famously articulated a distinction between *character* and *content* to properly account for indexicals such as "he"; character is the semantic property common to different uses of "he," content its contribution to what is asserted in well-behaved particular uses. I (2006) have argued that additional examples such as those provided before should be accounted for by generalizing the idea: Semantics for natural languages is "character-semantics." The semantics of a language as such does not identify the concrete speech acts that can be made with its sentences but merely constrains it; full determination of specific speech acts goes well beyond

what is provided by the language as such. In fact, this point applies also to the identification of the specific type of speech act that is made. The declarative mood of the whole sentence conventionally indicates that a speech act in the *saying* family is made; but whether it is one of *guessing, conjecturing, predicting,* or a default one of *asserting,* this depends on contextual considerations of "saliency"; the same applies to the other conventional indicators of speech act type such as the interrogative or the imperative moods.[7]

Speech acts such as assertion are thus essentially normative, and the meanings of natural language lexical items consist of their contributions to speech act potentials; the fact that we are sensitive to these two points is what, on the proposal I am outlining, our intuitive acceptance of claims such as (plus) and (circular) manifests. Similar remarks apply to the cognition case but in that case distinctions such as that between character and content are largely irrelevant.

25.3 Weak and Strong Normativity

In the previous section, I have outlined the initial considerations pointing in the direction of a normative account of meaning. Meanings, be they linguistic or cognitive, divide up (at least in fundamental cases) into a force and a representational content. The fact that (mental or linguistic) "lexical" items make a given contribution to representational contents has implications—of the kind (plus) and (circular) illustrate—for acts involving them; and the very nature of those acts themselves is to be understood in normative terms, along the lines of KR, TR, or perhaps RBR for the case of assertion.

What the considerations by Mulligan and Glüer and Wikforss mentioned in the introduction primarily show, I think, is that this is still a weak sense of normativity, which does not coincide with the strong one we are after—the one on which norms prescribe actions or provide reasons for agents to act. There is an easy way to appreciate this. Let us consider again the three rules we saw in the previous section that have been proposed to explain assertion, in the framework posited by Williamson, KR, TR, and RBR. It is common ground among the parties to this debate that assertion is what is done by default (i.e., unless conditions in an open-ended list apply, such as those creating irony, fiction, or the presence of canceling parenthetical remarks such as "I conjecture"—which allows a conjecture to be made by the utterance of a declarative sentence, etc) by uttering declarative sentences: "In natural language, the default use of declarative sentences is to make assertions," Williamson (1996/2000, p. 258). This gives us an independent, *causal-historical-*

[7] This might be in agreement with the difficult-to-interpret, disparaging remarks that Chomsky usually makes about truth and reference, cf. Pietroski (2003, 2006). I put "character" inside scare quotes to acknowledge Pietroski's claim that perhaps the meanings of natural language lexical items are not appropriately thought of as, strictly speaking, functions from context to semantic values; the term is merely used here to give a quick indication of the sort of conception of semantics I am gesturing towards.

intentional specification of the phenomenon that philosophers try to characterize: It is the act, whatever its proper characterization is, that is in fact associated with the indicative mood in natural languages as used on some occasions (the default ones), and which speakers intentionally purport to make by such means on such occasions. What is disputed is which of these norms an assertor is thereby subject to when uttering a declarative sentence in a default case. Of course, she may not be subject to any of them: After all, assertion might turn out not to be constitutively normative at all, as also indicated in the previous section, or might be characterized by a different norm altogether, or perhaps it only admits of a messy, disjunctive characterization appealing in part to some of those norms.[8]

Now, the obligations imposed by (TR) and (RBR), being constitutive of some act or other (even if it is not *assertion,* as causal-historical-intentionally picked out) "exist" in actuality (and in any other possible world, for that matter) as much as the obligation imposed by Williamson's (KR). To avoid confusions, we may call "t-assertion" and "rb-assertion" the acts defined by (TR) and (RBR), respectively, so as not to prejudice the issue which one of the two, if either, is assertion, causal-intentionally characterized. Similarly, I will call henceforth "k-assertion" the one defined by (KR). In these terms, the debate confronting Williamson, Weiner, and Lackey is whether assertion *is* k-assertion, t-assertion, or rb-assertion, if it is any of them at all.

All these acts are normative in at least this weak sense: They are types constitutively defined by obligations or permissions. This is normativity only in a *weak* sense because assertors are in actuality, at most, subject to one of the obligations imposed by the three purported rules we have considered. The others are perhaps alternative speech acts, causal-historical-intentionally specified under a different label—say, the (less determinate) act of *saying* in the case of (TR)—or perhaps those practices are simply "not in place" in the actual world, in fact actually committing nobody. Thus, t-assertion, rb-assertion and also k-assertion are not *strongly* normative, in the following sense: Their (Platonic) existence does not, by itself, give any actual rational being a (normative) reason to act.

The point is usefully made relative to games, the model on which constitutive-norms accounts of assertion are based. All possible constitutive norms for games "exist" in all possible worlds, on the assumption that games are governed by constitutive norms defining them. But not all of them govern the particular transactions of a group of individuals. As in the case of assertion, it might even be an epistemic achievement of sorts to determine which rules apply to given cases—which one among several different games, governed by slightly different sets of rules, is the one in fact being played. There is a causal-historical-intentional sense in which players are playing one and the same game *G,* while it might be unclear what the constitutive rules defining *G* are. To illustrate: Imagine a group of people playing a

[8] For skepticism about normative accounts, cf. Levin (2008); as indicated in the previous section, Hindriks (2007) defends Bach and Harnish's (1979) Gricean psychological, intention-based account, arguing that the norms applying to assertion are not constitutive but rather merely regulative, deriving from an independent, moral sincerity norm.

given card game, causal-historical-intentionally identified. They call it by the same name, "Rummy"; they have sufficiently similar dispositions and expectations with respect to the rules applying to it, which nobody has in fact set out to articulate explicitly and precisely. A dispute then arises regarding a specific rule of such a game, which they all call "Melding." Some players take the rule to be (BRM), others (HRM):

> (BRM) If you have a valid group or sequence in your hand, you may lay one such combination face up on the table in front of you. You cannot meld more than one combination in a turn.
> (HRM) If you have a valid group or sequence in your hand, you may lay one such combination face up on the table in front of you. You may lay as many such combinations in a turn as you want.

The description of the situation that the constitutive-rules account of games recommends goes as follows. There are two different games, which we may call BR-Rummy and HR-Rummy. If we use the term "Rummy," the causal-historical-intentionally individuated game that our players are in fact engaged in, then the dispute concerns whether Rummy is BR-Rummy or rather HR-Rummy (or, of course, perhaps neither of them, perhaps it is in fact some other game or perhaps the matter is just indeterminate). The same applies to assertion, I think: Assuming that assertion is, like games, individuated by constitutive rules, it is unclear (up for grabs, in the philosophical arena) whether assertion is k-assertion, t-assertion or rb-assertion, something else, or whether the matter is simply indeterminate.

By themselves, in sum, kinds characterized by means of constitutive rules are not strongly normative. We may presuppose that the kinds defined by constitutive rules are Platonic entities, existing in all possible worlds. This is a reasonable presumption, useful at least for expository purposes.[9] By way of analogy, Williamson (1996/2000, p. 239) mentions Lewis's (1975) proposal (in the context of a not unrelated discussion, as it will transpire in the next section) to conceive in this way the languages linguists characterize in a mathematically sophisticated way. On this view, any such characterization specifies a language, which, on the Platonist presumptions we are adopting, should not be called a *possible* language, because it "exists" in all the worlds, including the actual one (Schiffer 1993). But only some of those are actually *used by a given population*; there is a further issue about what makes it the case that a language is actually used by a population, an issue which might as well depend on contingent matters of fact.[10] Similarly, among all possible games and, in general, all kinds defined by constitutive rules, only some of them are actually "in force," giving actual people normative reasons to act—obligations and permissions.

[9] It can be argued that it is just an instrumentally convenient device, which can be later deflated of any excessive ontological implications by invoking some fictionalist strategy.

[10] I say "might turn out" instead of something stronger for reasons that will be elaborated in the next section. As I will say there, whether a language is in fact spoken—or a game played—by a given population turns on arbitrary conventional facts, and is thus a contingent matter. But it is doubtful whether this applies to promising and asserting. In those cases, it might well be that rationality is ultimately the source of the existence of the practices, which might be argued to be then a matter of (some sort of) necessity.

By itself, thus, Williamson's (KR), exactly like (TR) and (RBR), or the norms constitutive of different games are only weakly normative; only together with the assumption that k-assertion is what we in fact do when we assert (in the causal-historical-intentional sense) do they become strongly normative. By referring by "assertion" to the practice that the constitutive rule (KR) defines, we hide from ourselves the fact that, in the stronger sense, it is non-normative; this is because we pack into the characterization the fact that the practice so-defined is in fact implemented—it is the one people subject themselves to when they utter declarative sentences in default situations. But this is *additional* to the characterization provided by (KR).

Williamson is sensitive to the distinction I have made between the non-invidious Platonic existence of the types defined by constitutive norms, and their being in place, providing normative reasons to act to members of a given population, as his introductory clarifications make clear:

> Given a game G, one can ask, "What are the rules of G?" Given an answer, one can ask the more ambitious question "What are noncircular necessary and sufficient conditions for a population to play a game with those rules?" Competent unphilosophical umpires know the answer to the former question but not to the latter. Given a language L, one can ask, "What are the rules of L?" Given an answer, one can ask, "What are noncircular necessary and sufficient conditions for a population to speak a language with those rules?" Given a speech act A, one can ask, "What are the rules of A?" Given an answer, one can ask, "What are noncircular necessary and sufficient conditions for a population to perform a speech act with those rules?" […] assertion is presented to us in the first instance as a speech act that we perform, whose rules are not obvious. In order to test the hypothesis that a given rule is a rule of assertion, we need some idea of the conditions for a population to perform a speech act with that rule; otherwise we could not tell whether we satisfy those conditions. […] Our task is like that of articulating for the first time the rules of a traditional game that we play. (Williamson, 1996/2000, pp. 239–40)

In the next section, I will outline a tentative answer to the more ambitious question of the two Williamson contemplates here, what makes it the case that a given population plays a game, speaks a language, or performs a speech act, in all three cases on the assumption of an answer to the first question such that those are abstract entities consisting of the constitutive rules.

25.4 Institutional Practices

Rawls (1955) distinguished between *constitutive* and *regulative* norms; his aim was to vindicate rule-utilitarianism as opposed to act-utilitarianism, by making this "logical" or conceptual distinction. Take the obligations ensuing from promises. If promising is a constitutively natural, nonnormative activity, the obligation of keeping promises can be understood as a *regulative norm* on utilitarian grounds, relative

to the benefits it regularly confers.[11] Then it makes sense to consider, in the case of a particular valid promise, whether or not those benefits, and therefore the obligation, exist. Most of us find this intuitively wrong, as Rawls points out. This, he suggests, is because we think of the obligation of keeping promises as constitutive, definitional of the practice; if a valid promise has been made, then one must keep it.[12] This is a prima facie obligation, and hence could be overridden by other obligations with more force, including the utilitarian considerations; but even when it is, the prima facie promissory obligation still was in place.[13]

Even when the obligations related to promises are understood, as Rawls recommends, as constitutive or definitional, there is still a place for utilitarian considerations, he points out, now addressed to establish why the institution of promising, thus understood, should exist, be in place or implemented. Here Rawls is, I think, considering the same distinction I have made in the previous section, and pointing to the appropriate place at which, say, conventionalist claims may have a grip; later, in his main work (1972, pp. 344–350), he takes up the issue and gives a nonconventionalist account (unfortunately, called "conventionalist" in the literature on these matters, for reasons to be explained in a moment), as I will briefly explain now.

Consider again for the sake of comparison the case of games. Imagine that a group is playing the causal-historical-intentionally individuated game *G* but that there is an issue concerning which of the two possible sets of constitutive rules, Γ_1 and Γ_2, properly define *G*. As we have seen, the wide-scope obligations of which Γ_1 and Γ_2 consist "apply" to the situation, because they "exist" everywhere. This notwithstanding, as we have seen, in the strongly normative sense we have distin-

[11] This is the way Hindriks (2007) thinks of assertion, as we saw above. If I understand them well, Scanlon (1998, pp. 295–327, 2003) and Shiffrin (2008) provide, for the case of promises, contemporary defenses of the view that Rawls was criticizing, without calling upon utilitarian considerations; Scanlon (1998, p. 296; cf. 2003, p. 236) appeals instead to "a general family of moral wrongs which are concerned…with what we owe to other people when we have led them to form expectations about our future conduct," while Shiffrin (2008, p. 485) appeals to duties upholding an "ability to engage in special relationships in a morally good way, under conditions of equal respect." Owens' (2006, p. 51) is a third account; promises exist because of "what might be called an authority interest: I often want it to be the case that I, rather than you, have the authority to determine what you do," but he is noncommittal with respect to the issue of whether keeping promises is, in Hume's terms, a "natural virtue" such as beneficence—as Scanlon, Shiffrin, and the act-utilitarian with whom Rawls was debating would have it—or rather an "artificial" one, as on Rawls' (and Hume's) view, one dependent on the being in place of a practice consisting of a system of constitutive rules.

[12] Scanlon (2003, p. 245) is sensitive to this objection, and tries to specify a general principle that establishes normative obligations that exist whenever, intuitively, promises are valid, irrespective of whether more general moral obligations are overridden. He then argues that his principle "is not the social institution of promising under a different name" (2003, p. 247), but I am not convinced by his considerations.

[13] Moral considerations are still relevant, even assuming the constitutive rules account. Firstly, given their role in explaining why the institutions of promising exist—to be described below—they can help us to clarify which specific promising institution is it that we have in fact adopted, in particular the circumstances under which an act counts as a valid promise (Owens 2007). Secondly, they are relevant in particular cases to determine whether, all things considered, we should keep a promise.

guished, at most one of those sets of obligations is in place in the situation; disputes frequently arise among players about such matters, as in the example of Rummy we provided before for the sake of illustration. What makes it the case that, say, Γ_1 but not Γ_2 is in place in the situation (so that perhaps Γ_2 is not at all in place in the actual world)? It is in answering questions like this (the second kind of issue that Williamson mentions in the quotation at the end of the preceding section) that serious debates between conventionalists and their opponents have a grip.

In the case of games, a conventionalist account seems very plausible to me, but how should we properly articulate it? Here, it is useful to go back to Lewis' distinction between (abstract) languages and the language used by a given population. There is a way of thinking of abstract languages on which we do not have here just a useful analogy but one more example of the very same issue we are addressing: Alston's (2000) view of abstract languages as speech-act potentials, on the assumption that speech acts are characterized in terms of norms. An abstract language can be thought of as a pairing of sentences and appropriate meanings; as sketched before, we can think of those meanings as a constraint on the force and propositional content expressed, with the help of context, in a concrete literal, serious utterance of the sentence. In the case of a declarative sentence such as "I am hungry," the meaning assigned in English could be an indication of the rule to which assertors subject their act in default conditions, plus a Kaplanian character. What makes it the case that a language in which the sentence acquires this meaning is spoken by a given population, and hence that the constitutive rules partially defining it are in force in that population? Lewis' (1975) well-known answer is that a convention of using the language exists in that population.

Conventions, according to Lewis' (1969) account, are regularities that help to solve coordination problems, and are maintained because it is common knowledge that they have served this purpose in the past; they are arbitrary in that some other practice might have solved the coordination problem with a similar efficiency. Lewis' is, for the case of conventions, a reductive, "regulative norm" account of the kind Rawls was criticizing for the case of promises, and I would reject it for Rawls' sort of reasons; on the view I prefer, conventions should also be understood as non-reductively characterized by constitutive rules.

However, it is the existence of the kind of regularities that Lewis appeals to which is relevant to decide which practices defined by constitutive rules are in place, because, on the view I would like to suggest, this "being in place" of conventions supervenes on them. This is in fact the role Lewis (1975) assigns them in determining which among all possible languages—understood as abstract entities "existing" in all possible worlds—are in fact used or "exist" (in the other sense) at the actual world, as concrete tools of a given community. On the present proposal, conventions as much as languages themselves should be understood as defined by constitutive rules; Lewis' account explains which of the many conventions thus understood along Platonic lines, (DL) and (DR) among them, are actually in place, solving coordination problems in actual populations.

When it comes to other kinds defined by constitutive rules, like games, a conventionalist account contends on the present proposal that they exist if and only if a cor-

responding Lewisian convention implementing it exists. It is plausible, I think, to appeal to this kind of regularity to determine which set of constitutive rules, if any, defines *G* (without excluding, of course, the possibility that the issue is indeterminate, as I am sure it is in many actual cases). In accordance with Lewis' explanation of conventions, we should examine the dispositions players of *G* have to act and think (including in particular, as Alston (2000, pp. 265–8) insists, their normatively more relevant dispositions, such as to feel resentment or guilt, to apologize and criticize, and so on),[14] to settle the matter whether they conform their practice to the system of rules Γ_1 or rather Γ_2. On this proposal, a kind defined by constitutive rules is conventional if its existence (in the non-Platonic, invidious sense) is in fact the existence (in that very sense) of a Lewisian convention.

The conventionalist proposal about games, suggested here, is not simply the trivial one that certain means signaling that one is playing a given game (the figures on the cards that are used, the expressions players use in given moves) are conventional, in Lewis' sense; indeed they are, in the sense just explained: They are Lewisian conventions in fact implementing constitutive-rules conventions specifying the normative commitments incurred in using the relevant expressions. But the envisaged conventionalist claim goes beyond this; it is rather that the very fact that the constitutive rules of a given game actually apply in a given situation, their "being in place," is explained on the basis of what in fact constitutes the existence of a convention.

Rawls (1972, pp. 344–350), following Hume, is said to provide a "conventionalist" account of the promises, which other writers such as Scanlon (1998, 2003) and Shiffrin (2008) reject. However, Rawls' account is, I think, only "conventionalist" in the weaker sense that assumes the existence of a conventional practice of using certain expressions to indicate that a promise is being made—say, Lewisian (1975) conventions of truthfulness and trust in English regarding utterances of *S, I promise*, or *I hereby promise that S* in default situations.[15] In the case of promises, this almost non-contentious form of conventionalism is not questioned, say, by Scanlon or Shiffrin, who explicitly accept the existence of conventions of promising (better put, conventions of expressing promises) thus understood; but this is merely a proper part of the stronger form of conventionalism illustrated with the case of games. In that stronger sense, Rawls' (and Hume's) account is not, I think, conventionalist.

The reason is this. Lewis (1969) distinguishes from conventions what he calls *social contracts*. These are also regularities serving coordination problems kept in place because their previous existence is common knowledge, but they differ from conventions mainly in that, while in the latter case agents do not have any motivation for free riding, they do in the case of social contracts. A related additional

[14] Mulligan (1999, § 3) also mentions these psychological manifestations of Sprachgefühl, which he takes to concern rule-breaking and only in this way, indirectly, rule-following.

[15] A correspondingly weak conventionalism (that there are conventional means to express truths) is also trivial for the case of content; the philosophically interesting issue is whether or not there are truths made by convention, not just because they are claims about conventions that require conventions as truth-makers (Cf. García-Carpintero and Pérez-Otero 2009).

difference is that it is unclear whether they are arbitrary—whether there is another equally serviceable practice to solve the relevant coordination problem. In the case of social contracts, the explanation of conformity, and therefore of the preservation of the regularity, requires more than mere awareness of the existence of the practice; it requires sensitivity to moral norms, or other forms of commitments. Now, while driving on the right is a convention in Lewis' sense, it seems to me that promising and asserting are rather *social contracts*. Correspondingly, the explanation that Rawls (1972, pp. 344–50) provides for the "being in place" of the institution of promising is fundamentally moral; he appeals to sensitivity to a *principle of fairness,* requiring us to play our specified parts in social institutions which are fair and from which we benefit.[16]

I thus propose to classify views on these matters along the following lines. Promises and assertions are subject to norms but we have, firstly, the divide between what we might call *naturalists* and *institutionalists* accounts of this fact. Naturalists (Hindriks (2007) for assertion, and, I guess, Scanlon (1998) and Shiffrin (2008) for promises) think that these acts are not constitutively normative; they might be characterized in terms of Gricean reflexive intentions. Although we do have conventional practices indicating that they are performed, they can occur in their absence; both when they are made by relying on conventions, and in other cases, they are subject to general moral rules, perhaps duties of sincerity (Hindriks), duties to meet created expectations (Scanlon), duties to uphold intimate relationships (Shiffrin), or to promote "authority interest" (Owens). These are the sources of the more distinctive regulative norms applying to promises and assertions.

Institutionalists reject these views, mostly on the basis of variations on Rawls' argument against the utilitarian regulative norm account: The obligations resulting from a valid assertion or a valid promise still exist in cases in which, if we just took into consideration the moral benefits mentioned in naturalists accounts, they would not. They might be *overridden* by those moral considerations, so that agents do not have an *all-things-considered* obligation, but they still have the relevant prima facie obligation, as indicated by our intuitions regarding feelings of resentment or guilt or needs for excuses, apologies, compensation, and so on, which people do experience (Gilbert 2008, pp. 223–34). According to institutionalists, promises and assertions are defined by specific constitutive norms, norms that exist, are in place or have been adopted and therefore are strongly normative, in fact prima facie obliging actual people in actual situations.

This "being in place" consists in the existence of regularities, in their turn accounted for by specific dispositions to think and act in rational beings. But only

[16] The norms defining the institution of promising are, I think, more complex than the simple ones Williamson's account of assertion assumes, and this is what Rawls, following Searle, presupposes. Consequently, in the same way that (TR), (KR), and (RBR) specify different "assertion-like" practices, there are many different "promise-like" practices that we can specify; in particular, we can play with the "conceptual conditions" specifying further which specific circumstances of coercion, deception, etc. prevent a valid promise from being made. Only some of those practices would be, according to Rawls, fair, i.e., consistent with the two principles of justice that would be adopted in the original position (op. cit., p. 345).

some institutionalists are true *conventionalist*. The issue turns, I have suggested, on whether the regularities in question constitute a Lewisian convention, as opposed to a *social contract*; and hence, on whether or not we need to appeal to what is in fact sensitivity to moral norms (even if ultimately consisting themselves of norms of prudence or rationality, depending on which metaphysics of moral norms is correct) to explain their preservation. It could also be that the correct account combines elements from conventionalism and anti-conventionalism (both of them of the institutionalist variety): Promises and assertions described in an abstract, sufficiently general way are not conventional but when we go into the specifics of the particular practices that a given community has implemented we do find conventional features. I myself would endorse an institutionalist nonconventionalist account of assertion, promises, and (paradoxically as it might sound) conventions and agreements themselves, but this is not the place to examine the issue in any depth.

25.5 Conclusion

Let us take stock. Meanings, we have assumed, both linguistic and cognitive, consist in the fundamental cases of judgments, assertions, intentions, or enjoinments, individuated by a force and a representational content. The fact that (mental or linguistic) "lexical" items make a given contribution to representational contents, and in general the "truth-conditions" determined by those contents, has nothing normative in itself but it has implications—of the kind (Plus) and (Circ) illustrate—for acts involving them, whose very nature is to be understood in normative terms, along the lines of KR, TR, or perhaps RBR for the case of assertion, and corresponding norms for the case of argumentative transitions such as inductive or deductive inferences among belief-like states, or transitions from experiences to beliefs, etc. Again, the fact that the constitutive nature of those forces is normative does not by itself have strongly normative implications: It does not suffice to create reasons for agents to act in specific ways. However, together with whatever psychological or sociological facts determine that a given norm is in place, they do provide such reasons.

Let us then see how this answers the Glüer and Wikforss concern. Let us suppose that some acts of judging or asserting ("applying" a—perhaps mental—"term") are constitutive of a "term" having a certain meaning, for whatever reason.[17] In the case of the logical constants, such as the conditional, Carroll's (1895) famous paradox establishes that it cannot be judgments that a certain proposition follows from certain others, but inferential transitions relating them which are thus constitutive. It has not been sufficiently explored which norms could characterize such acts; for the sake of having something illustrative in mind, we can think of one corresponding to TR—a norm of conditional truth: One must ((infer p from the set $\sum$) only if

[17] The most obvious one would be that this is constitutive of the relevant meanings, along the lines of conceptual-role accounts, but, as Glüer and Wikforss point out, this could result from less controversial assumptions too.

(*p* is true on condition that all *q* in ∑ are also true)). Let us further assume that the relevant psychological or sociological facts establish that these norms are in place among us: TR for assertions/judgments, the norm just mentioned for deductive inferential transitions.

Now, are those assumptions compatible with the intuitively correct point that "oughts not only imply cans, they also imply the possibility of *violation*, of… 'forbidden combinations'"? Glüer and Wikforss are right, I think, that if applying "circular" to circular items in paradigm cases is determined as correct by the meaning of the term, and the corresponding point concerning the meaning of "if…then _," the cases we can think of in which these "norms" are "violated" do not count as proper "forbidden combinations"; they should have to be accounted for by facts such as lack of attention, excessive complexity, etc. However, the relevant norms we have been contemplating are norms for the representational acts (judgments, assertions, or inferential transitions) in general, not just in these particular cases linked to the meaning of some terms. There could then be many other cases allowing for proper violations: Cases that are not in any way constitutive of any of the occurring terms. And I think this is quite enough to satisfy the intuition that "forbidden combinations" should be allowed. Let us go on with the assumption that it is TR that is "in place" for our assertions. Should this intuitively require that wrongly asserting *that 2 + 1 does not equal 3,* say, is a proper "forbidden combination"? My own intuitions at least do not require that much. It requires that there are assertions involving those terms that do constitute forbidden combinations; but, of course, there are plenty of them.

This reply assumes that some explanation of which rules are "in place" along the lines of the one in terms of Lewisian conventions or social contracts suggested in the previous section can be properly developed: In effect, it assumes that there is an account of when rules properly "guide" us, and not merely characterize regularities we follow. Such an account should be able to circumvent another objection that Glüer and Wikforss pose (2009a, p. 52 ff.), the "dilemma of regress or idleness." We have to distinguish genuinely following rules from merely conforming to regularities; intuitively, the former requires some kind of mental attitude of *acceptance* by the subjects involved, i.e., the psychological states that the Lewisian account contemplates; but the considerations motivating normative accounts of foundational matters extend to these psychological states. Does not this lead to a non-virtuous regress? It certainly points to the need to encompass a sensitivity to norms which is natural to characterize with the Wittgensteinian metaphor as "blind," as the best recent work on the rule-following considerations has shown (cf. Boghossian 2008; Wright 2007). The problem is how to make articulated sense of this while still making room for genuine rule-following as opposed to mere rule-conforming. I cannot do any justice to these issues here.[18]

[18] A line that I like appeals to the teleological, functional-kind-based directives studied by different writers (cf. Thomson 2008, Chap. 12)—although for the reasons discussed in footnote 15 above, I am sure she would discourage this application of her views—and specially Jarvis (forthcoming).

Acknowledgments Financial support for my work was provided by the DGI, Spanish Government, research project FFI2010–16049 and Consolider-Ingenio project CSD2009–00056; through the award *ICREA Academia* for excellence in research, 2008, funded by the Generalitat de Catalunya; and by the European Community's Seventh Framework Programme *FP7/2007–2013* under grant agreement no. 238128. Thanks to Teresa Marques for helpful discussion of some topics in this review, and to Michael Maudsley for the grammatical revision.

References

Alston WP (2000) Illocutionary acts & sentence meaning. Cornell University Press, Ithaca
Austin J (1962) How to do things with words. Clarendon Press, Oxford. (Second edition issued as an Oxford University Press paperback, 1989, to which page references are made.)
Bach K, Harnish RM (1979) Linguistic communication and speech acts. MIT Press, Cambridge
Boghossian P (2008) Epistemic rules. J Philos 105:472–500
Carroll L (1895) What the tortoise said to Achilles. Mind 4:278–80 (reprinted (1995) Mind 104:691–93)
Dummett M (1978) Truth. In: Truth and other enigmas. Harvard University Press, Cambridge, pp 1–24
García-Carpintero, M (2004) Assertion and the semantics of force-markers. In: Bianchi C (ed) The semantics/pragmatics distinction. The University of Chicago Press, Chicago, CSLI Lecture Notes, pp 133–166
García-Carpintero M (2006) Recanati on the semantics/pragmatics distinction. Crítica 38:35–68.
García-Carpintero M, Pérez-Otero M (2009) The conventional and the analytic. Philos Phenomen Res 78:239–274
Gilbert M (2008) A theory of political obligation. Oxford University Press, Oxford
Glüer K, Wikforss Ã (2009a) Against content normativity. Mind 118:31–70
Glüer K, Wikforss Ã (2009b) The normativity of meaning and content. The Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/archives/sum2009/entries/meaning-normativity/. Accessed 17 June 2009
Green M (2007) Self-Expression. Oxford University Press, Oxford
Grice HP (1975) Logic and conversation. In: Cole P, Morgan J (eds) Syntax and semantics, vol 3. Academic Press, New York
Hindriks F (2007) The status of the knowledge account of assertion. Linguist Philos 30:393–406
Jarvis B (forthcoming) Norms of intentionality: norms that don't guide. Phil Stud 157:1–25
Kaplan D (1989) Demonstratives. In: Almog J, Perry J, Wettstein H (eds) Themes from Kaplan. Oxford University Press, Oxford, pp 481–563
Kripke S (1982) Wittgenstein on rules and private languages. Harvard University Press, Cambridge
Lackey J (2007) Norms of 0020Assertion. Noûs 41(4):594–626
Levin J (2008) Assertion, practical reason, and pragmatic theories of knowledge. Philos Phenomen Res 76(2):359–384
Lewis D (1969) Convention: A Philosophical Study-Cambridge, Mass.: Harvard University Press
Lewis D (1975) Languages and language. In: Gunderson K (ed) Language, mind and knowledge. University of Minnesota Press, Minnesota, pp 3–35. (Reprinted in Lewis, D (1983) Philosophical Papers, vol 1. Oxford, Oxford University Press, pp 163–188)
Mulligan K (1987) Promising and other social acts: their constituents and structure. In Mulligan K (ed) Speech act and sachverhalt: reinach and the foundations of realist phenomenology. Nijhoff, Dordrecht, pp 29–90
Mulligan K (1999) Justification, Rule-Breaking and the Mind. Proc Aristotelian Soc 99:123–139
Owens D (2006) A simple theory of promising. Phil Rev 115:51–77
Owens D (2007) Duress, deception and the validity of promises. Mind 116:293–315

Peacocke C (1987) Understanding logical constants: a realist's account. Proc Brit Acad 73:153–200
Pietroski P (2003) The character of natural language semantics. In: Barber A (ed) Epistemology of language. Oxford University Press, Oxford, pp 217–256
Pietroski P (2006) Character before content. In: Thomson J, Byrne A (eds) Content and modality: themes from the philosophy of Robert Stalnaker. Clarendon Press, Oxford, pp 34–60
Rawls J (1955) Two concepts of rules. Phil Rev 64:3–32
Rawls J (1972) A theory of justice. Oxford University Press, Oxford
Scanlon TM (1998) What we owe to each other. Harvard University Press, Cambridge
Scanlon TM (2003) Promises and contracts. In: Scanlon TM (ed) The difficulty of tolerance. Cambridge University Press, Cambridge, pp 234–269
Schiffer S (1993) Actual-language relations. In: Tomberlin J (ed) Philosophical perspectives 7: language and logic. Ridgeview, Atascadero, pp 231–258
Searle J (1969) Speech acts. Cambridge University Press, Cambridge
Shiffrin S (2008) Promising, intimate relationships and conventionalism. Phil Rev 117:481–524
Thomson JJ (2008) Normativity. Open Court, Chicago
Vlach F (1981) Speaker's meaning. Linguist Philos 4:359–391
Weiner M (2005) Must we know what we say? Phil Rev 114:227–251
Williamson T (1996/2000) Knowing and asserting. Phil Rev 105:489–523 (reprinted with some revisions as chapter 11 of: Williamson, T (2000) Knowledge and Its Limits. New York, Oxford University Press, from which I quote)
Wright C (2007) Rule-following without reasons: Wittgenstein's Quietism and the constitutive question. Ratio 20(4):481–502

Chapter 26
M&Ms—Mentally Mediated Meanings

Laurent Cesalli

Abstract Aristotle's sketchy remarks on the relations between words, concepts and things at the beginning of the *De interpretatione* (16a3–9) gave rise to several competing interpretations during the Middle Ages. All of them claimed to be orthodox, not only with respect to the Stagirite himself, but also to Boethius, the first and most significant Latin authority on that matter. The mainstream interpretation (Peter Abelard, Thomas Aquinas, Walter Burley, John Duns Scotus, John Buridan) takes Aristotle to mean that words signify things by means of concepts in the sense that words signify concepts immediately, and things only mediately—an idea which is captured in the leitmotiv *voces significant res mediantibus conceptibus*. From the middle of the thirteenth century onwards, Aristotle's (and Boethius') words also gave rise to competing, sometimes quite idiosyncratic, interpretations of the exact role played by concepts in the semantics of words. These interpretations share the rejection of conceptual mediation and claim either that words signify things conventionally and concepts naturally (Roger Bacon), or that words do not signify concepts at all and things in a derivative way only, namely insofar as words signify conventionally the very same things which concepts signify naturally (William of Ockham). Obviously, the discussion of the *mediantibus conceptibus* thesis did not end with the Middle Ages, for one can trace its continuous use and discussion through the Second Scholasticism—mainly in seventeenth century Spain—and subsequently in the vast pedagogical literature aimed at providing students in general (and theologians in particular) with a "classical" (i.e. scholastic) philosophical background. It is not surprising, then, that a late nineteenth and early twentieth century philosopher of language like Anton Marty uses the *mediantibus conceptibus* principle and claims that it also applies to his own semantics. Thereby, the Swiss philosopher does nothing but add a page—though a remarkable one—to the long history of the reception of Aristotelian semantics. My aim is neither to discuss Marty's or the medievals' interpretative models, nor to show how exactly the *mediantibus conceptibus* thesis was transmitted from the seventeenth to the nineteenth century. Rather, I intend to gather some first indications in order to answer the following question: Which are the doctrinal similarities and differences between Marty's understanding of the

L. Cesalli (✉)
CNRS, Université de Lille 3 (UMR 8163: Savoirs, Textes, Langage)
19b, rue des Champs du Bourg CH-1920 Martigny, Switzerland
e-mail: lcesalli@gmail.com

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_26,

mediantibus conceptibus thesis and the philosophical tradition which obviously—though indirectly—inspired it?

Keywords Philosophy of Language · Philosophy of Mind · Medieval Philosophy · Anton Marty

26.1 Anton Marty and the *Mediantibus Conceptibus* Principle

On at least six occasions, Anton Marty (Schwyz 1847–Prag 1914) mentions the scholastic saying according to which words signify things by means of concepts: *Voces significant res mediantibus conceptibus* (on Marty's philosophy, see Mulligan 1990; Baumgartner et al. 2006/2009; Rollinger 2010). The general context of these occurrences is the semantics of names or, in Marty's terminology, of *Vorstellungssuggestive*. The claims, in whose defence Marty alludes to the medieval principle, are the following:

1. In the sentence "*A* exists", *A* does not name a concept[1] but an object. This, says Marty, was the opinion of the (good) scholastics:[2]

   > Die Namen, sagte sie [i.e. die bessere Scholastik], *bezeichnen die Dinge*; doch tun sie es unter Vermittlung der Begriffe (*mediantibus conceptibus*). Daher gibt es allgemeine und individuelle Namen, wie es allgemeine und individuelle Begriffe gibt. (Marty 1894/1918, p. 168, note 1)

2. Strict synonymy occurs when different names name the same object by means of the same concept (a variation of the mediating concept excludes synonymy, but a variation of the auxiliary concepts, which ensure the link between the word and its meaning is neutral with respect to synonymy; Marty calls such auxiliary concepts the inner linguistic form):

   > In Wahrheit haben die Vertreter der aristotelischen Logik trotz ihres Satzes: '*Vocabula sunt notae rerum*' nicht verkannt, dass die Namen in gewissem Sinne auch Zeichen unserer Begriffe sind. Ja, es ist sogar ihre Lehre, dass sie letzteres mehr direkt und ersteres nur indirekt und mittelbar sind. Die Namen—darüber war die aristotelische Logik sich völlig klar—können Zeichen von etwas in mehrfachem, namentlich in doppeltem Sinne genannt werden, indem sie es *bedeuten* oder indem sie es *nennen*. Letztere Funktion ist vermittelt, und zwar durch die erstere. Die Namen sind Zeichen unserer Begriffe oder Vorstellungen, indem sie solche in uns erwecken. Das Aussprechen eines Namens ist ein Mittel, im Hörer

[1] Here, "concept" renders "*Begriff*"—and just as a concept is a conceptual presentation (*begriffliche Vorstellung*), a perception (*Anschauung*) is a "sensible" presentation (*anschauliche Vorstellung*). In the following (and when considering Marty's position), we take "concept" in the sense of a conceptual presentation.

[2] Marty (1894/1918, p. 168): "In dem Satze 'A ist' nennt A nicht einen allgemeinen oder individuellen Begriff, nicht den Begriff eines Kreises oder den Begriff dieses Buches, sondern wie schon die bessere Scholastik betont hat, einen Kreis oder dieses Buch selbst, und das ist es, was in dem ausgesprochenen Urteil anerkannt wird".

einen gewissen Begriff hervorzurufen, und ihn nennt man darum die Bedeutung oder den Sinn des Namens. Ein Lautkomplex, der keinen Begriff erweckt, ist für uns 'sinnlos'; solche, die denselben Begriff wachrufen, heissen gleichbedeutend oder synonym. Fragt man jedoch, was der Name nenne, so ist es nicht der Begriff oder die Vorstellung, sondern der Gegenstand derselben, das, was ihnen etwa in Wirklichkeit entspricht. Aber nur *mediantibus conceptibus*, wie die alte Logik richtig sagte, werden die Gegenstände durch die Laute unserer Sprache genannt, *unter Vermittlung der Begriffe* und als das, als was die Begriffe sie auffassen. Diesen und keinen andern Sinn hatte der Satz: '*vocabula sunt notae rerum*', und zu einem Tadel desselben ist für den, der ihn nicht missdeutet, kein Anlass. Die Tatsache, dass ein *Gegenstand* unter Vermittlung verschiedener—mehr oder weniger vollständiger—Begriffe eine mehrfache Benennung erhalten kann, ist also altbekannt. Neu ist ihre Vermengung mit der ganz andern, dass oft *derselbe Begriff* unter Vermittlung verschiedener innerer Formen einen lautlichen Ausdruck empfängt, und man kann nicht vorsichtig genug sein, sie zu vermeiden und auszuschliessen. (Marty 1893/1920, p. 84, note 1) [3]

3. The function of nomination (Nennung) depends on the functions of indication (or Kundgabe, i.e. the speaker's indication of a certain presentation of his) and steering (or *Bedeutung*, i.e. the triggering of a certain presentation in the hearer).[4] In a footnote, Marty comments:

   In diesem Sinne kann man es nur billigen, wenn schon die Scholastiker sagten: *Voces significant res mediantibus conceptibus*. Die Namen nennen in der Tat die Gegenstände als das, als was sie durch unsere begrifflichen Gedanken erfasst werden (resp. vom Hörer erfasst werden sollen). (Marty 1908, p. 436, note 1)

4. The medieval semantic principle holds for proper names as well: What they name is mediated by what they mean, but in their case, the mediating concept is a singular one, and the context of the utterance determines which concept plays the mediating role:

   Der Satz, dass die Namen die Gegenstände nennen *mediantibus conceptibus* könnte nur eine Anfechtung erfahren hinsichtlich der Eigennamen im engsten Sinne des Wortes, wie Aristoteles, Napoleon ... <Man> könnte geneigt sein zu meinen, die Eigennamen nennten bloss etwas, *ohne etwas zu bedeuten*. Doch wäre auch dies meines Erachtens nicht die richtige Deutung der Tatsachen. Auch hier wird eine Vorstellung des einzelnen Gegenstandes, die seine Nennung vermittelt (und natürlich muss es eine individuelle sein), nicht fehlen. Aber es ist dem Zusammenhang überlassen, *welche* gerade erweckt werde, während der Name für sich allein in dieser Beziehung nicht determinierend wirkt. (Marty 1908, pp. 438–439, note 2)

5. In the case of names expressing conceptual presentations (Begriffe, as opposed to Anschauungen), the scholastic principle can be understood in a notable way. The mediating function can be attributed not to the concept itself, but to its content—the content of the concept "white", for example, is that in virtue of which every white object falls under that concept. It is what Marty calls "the concept's

[3] Here Marty criticizes Noiré (1882, p. 368) and his quite idiosyncratic view of ancient logic: "Aristoteles hielt also die Namen für Symbole der Dinge (*vocabula sunt notae rerum,* Cic.). Um zu dem Begriffe des Begriffs (*notio, conceptus*) zu gelangen, bedurfte es nicht weniger, als der angestrengten Geistesarbeit der ganzen mittelalterlichen Philosophie". Noiré probably took his Ciceronian quotation from Herder's *Abhandlung über den Ursprung der Sprache* of 1772.

[4] Marty (1908, p. 436): "Unter Vermittlung dieser äussernden und jener Bedeutungsfunktion aber kommt den Namen nun auch das zu, was wir als das *Nennen* bezeichnen".

object in a strict sense" as opposed to "the concept's object in a wide sense", i.e. its extension:

> <Es> kann bei ihnen [i.e. den Begriffen], im Gegensatz zu den Anschauungen, ein Gegenstand im engeren und weiteren Sinne unterschieden werden. Im weiteren Sinne wäre also z.B. für den Begriff Weisses alles das ein Gegenstand zu nennen, was zu seinem Umfang, d.h. zum Bereiche seiner Anwendbarkeit gehört, d.h. alles, wovon, wenn es ist, das Weisssein in Wahrheit prädiziert werden kann. Im engeren Sinne dagegen kann vom Gegenstand dieses Begriffes gesprochen werden mit Rücksicht auf diejenige Seite an dem im ersten Sinne Gegenstand Genannten, wonach dieses in einer solchen unvollständigen Vorstellung erfasst ist. Mit anderen Worten: wenn das Vorgestellte, falls es wirklich wäre, dem Vorgestellten in der Art adäquat sein würde, daß in ihm nichts gegeben wäre, was nicht auch im Vorstellenden als solchen sein Gegenstück hätte, so können wir es 'Gegenstand im engeren Sinne' oder *Inhalt* nennen...Fasst man den Terminus 'Inhalt' in dieser Weise...so kann man, statt zu sagen: die Namen nennen die Dinge *mediantibus conceptibus* oder als das, als was sie vorgestellt werden, sich auch ausdrücken: sie bedeuten den Inhalt unserer begrifflichen Gedanken, die wir durch Aussprechen des Namens als in uns stattfindend äussern und in gleicher Weise im Hörer erwecken wollen. (Marty 1908, p. 448)

6. The difference between proper and improper presentations is a difference in mediating concepts—an improper presentation is a presentation which is not perceptual (neither directly nor indirectly), or which presents its object in an oblique and nonessential way (i.e. something like a case of a mental *denominatio extrinseca*):

> Es kommt bekanntlich vor, dass man etwas ausdrücklich as *eigentlich* 'unnennbar' bezeichnet, es aber doch irgendwie nennt; also *uneigentlich*. Was ist mit diesem Unterschied eines eigentlichen und uneigentlichen Nennens oder Bezeichnens gemeint? Eines dürfte klar sein, dass, wie überhaupt die Namen stets etwas nennen *mediantibus conceptibus*, jener Unterschied in der Weise des Nennens mit einem Unterschied der dasselbe vermittelnden 'Begriffe' zusammenhänge. (Marty 1908, p. 455)

Without going into the details of the many doctrinal elements appearing in the six passages just quoted, one can perhaps summarize Marty's understanding and use of the scholastics' semantic principle as follows: The "*mediantibus conceptibus*" encompasses several moments in the semantics of names whose careful distinction is probably Marty's most significant contribution to the philosophy of language. In a certain sense—but in a certain sense *only*—names are signs of concepts: Concepts are not the terms of the semantic relation, but the successful use of names *always* involves the mental level of presentations. Three "conceptual" moments can be distinguished in Marty's analysis of the meaning of names.[5] The first two are part of the complex pragmatic process of *Bedeutung:* A name is a vocal tool used to indicate a certain concept in the speaker (*Kundgabe*), and to trigger a certain

[5] The full analysis of communication involves the further level of the "inner linguistic form" (*innere Sprachform*). The presentations making up the inner form, however, are not part of what is meant by names. They are auxiliary presentations linking the uttered sounds to the intended presentations. Presentations making up the inner form are implicit (and often unnoticed) leftovers of earlier stages in the development of a given language, mainly functioning by habit and association of ideas (see Marty, 1908, pp. 134–150).

concept in the hearer[6]—a process which includes a genuine normative component, for in the proper sense of the expression, the *Bedeutung* of a name is "that the hearer *should* form a certain presentation (*Vorstellung*)"[7]; the third moment consists in the intentionality of the (indicated and triggered) presentation: A presentation always presents an object and that object is precisely what names name (*Nennung*) (on Marty's theory of intentionality, see Chrudzimsky 1999). Thus, a kind of mental (or conceptual) mediation is crucial in Marty's "pragmatic semantics":[8] Words refer to things by indicating and triggering mental acts directed towards objects.[9]

26.2 The Aristotelian–Boethian Framework

Concerning the question of Marty's source for his use of the medieval formula *voces significant res mediantibus conceptibus,* the Swiss philosopher could have found it just about anywhere in the vast pedagogic literature of his times.[10] Like his teacher and friend Franz Brentano, Marty was (for a surprisingly short time) a catholic priest, a circumstance which suggests a certain familiarity with compendia of scholastic philosophy, the kind of books on which the philosophical training of

[6] *Bedeutung* is itself mediated by *Kundgabe* (Marty 1894/1918, p. 69): "Aber...die Kundgabe meines Vorstellens bleibt doch stets das Mittel zur Erweckung der Vorstellung im Hörenden, und diese letztere Funktion des Namens, die wir seine Bedeutung nennen, kommt ihm also nur mittelbar zu vermöge der ersteren, die wir sein Kundgeben nannten. Der Name ist Zeichen einer Vorstellung, die der Hörende in sich erwecken soll, indem er Zeichen des Vorstellens ist, das im Redenden sich abspielt". For a qualification of this principle (in practice, one only communicates by means of sentences, and never by using pure, isolated names), see Marty (1908, p. 491).

[7] See for example Marty (1908, p. 288): "Indem wir die Begriffe des Bedeutens im allgemeinen erläuterten und speziell auch am Beispiel der Aussage illustrierten, sind wir bereits zu dem Resultate gekommen: die Bedeutung der Aussage sei es, im Hörer ein Urteil von bestimmter Art zu erwecken. Statt dessen kann man sich aber auch ausdrücken: die Aussage bedeute 'dass der Hörer ein gewisses Urteil fällen solle". —On that topic, see the monograph (inspired by Marty) of Erich Ahlman (1926).

[8] Calling the first two moments "semantic–pragmatic", one can qualify the third as "strictly semantic"—in that sense, Marty's semantics is a "pragmatic semantics".

[9] For an attempt to bring out the details of Kundgabe and Bedeutung with respect to the intentions these processes essentially involve, see Cesalli (2013).

[10] The *Cursus philosophici* of the main figures of the second scholasticism were still widely used at the end of the nineteenth century. See, for example, Cosmus Alamannus, Summa philosophiae ex variis libris D. Thomae Aquinatis Doctoris Angelici in ordinem cursus philosophicus accommodata, Parisius: Lethielleux, 1885 (1618 for the first edition), p. 221: "Respondeo [to the question: utrum nomina et verba immediate significant conceptus an res] dicendum, quod necesse est dicere, quod nomina et verba significant res et conceptus, sed hos immediate et per prius, illas vero mediate, mediantibus scilicet conceptibus, et per posterius". For (many) other occurrences of this leitmotiv in the philosophy of language of the sixteenth and seventeenth centuries, see Stephan Meier-Oeser, Die Spur des Zeichens. Das Zeichen und seine Funktion in der Philosophie des Mittelalters und der frühen Neuzeit, Berlin: De Gruyter, 1997, esp. 282 ff.

future ecclesiastics was based.[11] Whichever Marty's immediate source might be, the formula certainly developed from Boethius' (d. 524) comments on Aristotle's *De interpretatione*.[12] In the opening lines of this short treatise, Aristotle writes:

> Spoken expressions are symbols of mental impressions, and written expressions <are symbols> of spoken expressions. And just as not all men have the same writing, so not all men make the same vocal sounds, but the things of which <all> these are primarily signs are the same mental impressions for all men, and the things of which these <mental impressions> are likenesses are ultimately the same. (Aristotle 1980, *De interpretatione*, c. 1, 16a3-9, p. 30)[13]

Engaging in a kind of comment *avant la lettre,* Boethius writes, in the much longer *editio secunda* of his work:

> And so before coming to Aristotle's own words, let us discuss a little in general verbs and names and what they signify. For if there is to be questioning and answering or continuous and coherent speech so that another person hears and understands, if anyone is to teach, another learn, the whole arrangement of speech consists of these three: things, thoughts and spoken sounds. The thing is conceived in a thought. Spoken sound signifies the concepts of the mind and thoughts, whilst the thoughts themselves both conceptualise the things which underlie them and are signified by spoken sounds. (Boethius 2010, p. 25)

Two points can be underscored here: First, the different semantic relations present in Aristotle's text (being symbol, sign, or likeness) are reduced by Boethius to the sole relation of being a sign of something. Second, the basic Aristotelian scheme is explicitly placed in a communicational setting (questioning and answering, another person's understanding, teaching and learning)—something which, for sure, is not excluded by Aristotle's words, but on which he certainly does not put any emphasis.[14] In fact, as Magee (1989, pp. 64–92) has convincingly shown, the interplay between a speaker and a hearer plays a crucial role in Boethius' exposition, for he considers the triad *res-intellectus-vox* from two opposite directions, corresponding to the order of speaking (*ordo orandi,* from *vox* to *res*) and the order of knowledge (*ordo cosgnoscendi,* from *res* to *vox*):

> The man who teaches or gives a continuous address or asks questions behaves in the opposite way to those who learn, listen or reply, in three things: spoken sound, thought and thing (we will leave out the letters because some people cannot read). For those who teach, speak

[11] Marty completed his training in theology in 1867 in Mainz with a work on Aquinas' theory of knowledge. In the following year, he began attending Brentano's lectures in Würzburg, where he also met Carl Stumpf. His time as a priest lasted from 1870 to 1873 (Brentano, himself a priest since 1864, made the same radical move in the same year).

[12] On Boethius' theory of language, see the outstanding study of John Magee (1989). On Boethius' philosophy and its influence on the Middle Ages, see Marenbon (2003, 2009).

[13] Magee judges Apostle's translation "more precise" than Ackrill's version of the text.

[14] Note, however, that a little later in the *De interpretatione* (c. 3, 16b19–20), Aristotle says that "[w]hen uttered just by itself a verb is a name and signifies something—the speaker arrests his thought and the hearer pauses—but it does not yet signify whether it is or not". As we shall see, medieval commentators will associate this passage with Aristotle's theory of linguistic meaning at the beginning of the same treatise which gave rise to the very idea of the concepts' mediating role. And in Rhetoric I.3, 1358b1, he says that three components make up speech: the speaker, that which is talked about, and the hearer, the latter being the most important element.

> and question, proceeding from things to a thought, practise the function and power of their own particular activity through names and verbs. For they derive their thoughts from things which act as subjects and express them through names and verbs. But the man who learns, hears or even the man who answers, goes from names to thoughts and eventually reaches the things. The learner, listener or answerer, in receiving the words of the teacher, speaker or questioner, understands what each of them says and in understanding acquires knowledge of the things too and is confirmed in that knowledge. (Boethius 2010, p. 27)

In the resulting picture, concepts play a pivotal role in both orders. As Magee nicely puts it, they possess a "Janus like quality", being "in some sense 'contiguous' with both, *res* and *vox*" (Magee 1989, p. 71). Accordingly, they function as mediators in the cognitive order (one could say something like *res concipiuntur mediantibus conceptibus*), as well as in the corresponding (and symmetrical) semantic order: *Voces significant res mediantibus conceptibus*.

26.3 Four Medieval Models

The most important source for medieval semantic theory is Aristotle's isolated remark on the meaning of words at the beginning of the *De interpretatione*.[15] Boethius's detailed explanation of this laconic passage plays a crucial role in the reception of the Stagirite's thought. The phenomenon of linguistic meaning (*significatio*) is explained in cognitive *and* communicative terms: In the hearer, words lead to things via concepts *because* the words pronounced by the speaker express concepts of things.

Although the Aristotelian–Boethian consensual idea that concepts play an essential role in the phenomenon of linguistic meaning is traditionally cashed out by the *mediantibus conceptibus* thesis—i.e. by the claim that concepts are immediately signified, and things only through concepts—the exact role of concepts gave rise to what the medievals perceived as a *non modica contentio inter viros famosos,* in the words of Roger Bacon, or, as John Duns Scotus puts it, as a *magna altercatio*.[16] In the following, we shall consider four competing semantic models elaborated from the twelfth to the fourteenth century on the base of the Aristotelian–Boethian framework. It will turn out that the double dimension of *cognition* and *communication* already clearly identified by Boethius will lead to the emergence of two nonexclusive perspectives: a strictly semantic and a pragmatic–semantic perspective, the former being typically expressed by the *mediantibus conceptibus* thesis, and the latter by the idea that "to signify" (*significare*) is nothing but to constitute a concept in the mind of the hearer: *significare est intellectum constituere*.

---

[15] Other important sources are Augustine's *De magistro and De doctrina christiana*. The *De dialectica* had close to no impact on medieval semantics (Rosier-Catach 1995).

[16] On medieval semantics and its development, see for example De Rijk (1967); Pinborg (1971); Biard (1989); Rosier-Catach (1994); Jacobi et al. (1996); Meier-Oeser (1997, pp. 42–114, esp. 82–86); Pini (1999); Amerini (2000); Pini (2001); Valente (2008); Marmo (2010).

26.3.1 Concepts As Semantic Mediators

In his glosses on Aristotle's *De interpretatione*—a commentary he wrote on the base of Boethius' translation and (second) commentary—Peter Abelard (1079–1142) argues at length that, at least in standard cases, words first and foremost signify concepts (*intellectus*) which lead us to the cognition of things[17]:

> Vocal sounds were not invented because of the similitudes of things or of concepts, but rather because of the things themselves and their concepts, that is: in order that those concepts produce knowledge about the natures of things ... Vocal sounds ... constitute concepts about things..., since it is clear that vocal sounds direct [*applicant*] the hearer's mind on a similitude of a thing in order that in this <similitude> the mind focuses [*attendat*] not on <the similitude> itself, but on the thing for which <the similitude> was posited. (Abelard 2010, c. 1, p. 33–34)

This explanation is remarkably hearer-oriented, a perspective which is related to the general (and traditional) question of the utility of the treatise one is commenting upon. In the prologue of his commentary, Abelard, relying on the opinion of Herminos—a Greek commentator, active in the second century A.D.—makes the following observation:

> When Aristotle teaches that vocal sounds are the marks of concepts, he indicates the utility of the intended work. For since everyone is naturally apt to perceive concepts, it was useful to know by means of which instruments [*instrumenta*] one could manifest one's concepts or conceive foreign <concepts> [*suos intellectus manifestare uel alienos concipere*]—something Aristotle evidently suggests where he shows, in this treatise, that vocal sounds are marks of concepts. And therefore, the so to speak general and common utility of the whole treatise consists in teaching us that vocal sounds can produce [*generare*] or contain [*concipere*] concepts. (Abelard 2010, p. 21–22)

The communicative interpretation line opened by Boethius is obviously taken up by Abelard who insists on what we could call a "pragmatic moment" in his semantics. Linguistic meaning is not only *significatio,* a (mediated) relation between words and things, it is also *significare,* an action performed by a speaker with a precise goal—*significare est intellectum constituere* as Abelard (also) says in his treatise on concepts (Abelard 1994, § 91):

> And <verbs and names> "signify something", and that they do signify, <Aristotle> shows on the base of the description of what it is to signify [*a descriptione significandi*], namely: that they determine [*constituunt*] a concept in the hearer. And this is: he "who speaks", that is: who utters a word, "arrests his thought". Which he [i.e. Aristotle] shows from the effect, that is: because he "who hears" a word "pauses", namely: focuses [*haerendo*] and freezes [*figendo*] his mind in the concept which it has through the word. (Abelard 2010, c. 3, p. 116)

[17] Nonstandard cases discussed by Abelard are instances of empty terms (meaning, nonexistence or fictions). In such cases, concepts are signified, but there is nothing in the extra-mental world to "terminate" (i.e. to be the target term of) the relation of signification. The semantics of empty terms provides a strong argument for the immediate signification of concepts by words (Abelard 2010, pp. 28–39).

Thus, Abelard links the concepts' *mediating role* with the *instrumental function* of words: Linguistic meaning is explained in terms of speakers' linguistic actions and their effects on hearers.[18]

Writing a good century after Abelard, Thomas Aquinas (1224–1274) is the most famous medieval advocate of the Boethian reading of the opening chapter of the *De interpretatione*. In his comment, he puts forward a metaphysico-doxographical explanation for the thesis that words signify things *mediantibus conceptibus*: Since names have a general meaning—"*homo*" does not signify Plato rather than Socrates—and since the Stagirite rejected Plato's Ideas, words simply *have* to signify things via concepts:

> it is important to understand that the intellect's conceptions are what names and verbs signify, according to Aristotle: for it cannot be that they immediately signify things themselves, as is obvious from their mode of signifying: for this name 'man' signifies the human nature abstracted from singulars, and thus, it cannot be that it signifies immediately a singular man. This is the reason why the Platonists claimed that it signifies the separate idea of man itself; however, because, according to Aristotle's opinion, such <an idea> does not really subsist in abstraction, but only in the intellect, Aristotle had to say that words signify conceptions of the intellect immediately, and through them, things [*res mediantibus illis*]. (Aquinas 1989, p. 10–11)[19]

As for the pragmatic or instrumental dimension introduced by Abelard, we find it in Aquinas as well, though with lesser emphasis. Commenting on the thesis of the semantic equivalence of names and verbs (*De interpretatione* 3, 16b19–20), Aquinas says that the *proprium* of a signifying word is to evoke or produce a concept in the mind of a hearer: "*proprium vocis significativae est quod generet aliquem intellectum in animo audientis*" (Aquinas 1989, I.5, p. 29). Here again, the concepts' mediating role is not only stated, but also presented from the perspective of verbal communication.

[18] Such an interpretation seems to have been common from the second half of the twelfth century onward. See for example, the following passage of the anonymous Ars meliduna (written between 1160 and 1180), in Lambert Marie De Rijk, Logica modernorum, vol. II.1, p. 296: "Words have been imposed upon things not because of the things themselves—indeed, these do not require imposition for their manifestation, since they are offered to sense perception—but in order to interpret the concepts of things we have. However, it would be more appropriate to say that words are interpreted [*interpretari*] or that they constitute [*constituere*] concepts, than that they signify".

[19] The *mediantibus conceptibus* thesis is the *opinio communis* in the thirteenth century, defended for example by Robert Kilwardby, Albert the Great, and Lambert of Lagny (Marmo 2010, p. 115, with further literature). One should note, however, that different accounts of the nature of the signified and mediating mental entities are propounded. Thus, the three authors just mentioned identify them with the *species intelligibilis,* whereas for Aquinas it is the *verbum mentis* (or *conceptio intellectus*) which plays this role, a mental entity which depends on, but is distinct from, the *species intelligibilis* (Thomas Aquinas, *Quaestiones disputatae De potentia*, q. 8, a. 1, resp.).

26.3.2 Against Semantic Mediation: The Direct Signification of Things

The interest for what actually happens between users of a given language is even stronger in Roger Bacon (1214–1292), author of a recently discovered (and incomplete) treatise on semiotics and semantics entitled *De signis*.[20] Among other original views, Bacon holds that linguistic meaning is a dynamic property. Taking the conventionality of language at face value, he claims that speakers are free to, and in fact often do, (re)impose words in their actual linguistic interactions, thus using them as signs for variegated "things" (in the broadest sense of the word, that is, including concepts and nonexistent items) according to what they *mean to say* by those words.[21]

No formal commentary of Aristotle's *De interpretatione* by the Franciscan master is known, but at the end of the fragment of the *De signis* which has come down to us, Bacon addresses the question of the relation between words, concepts and things. In a certain sense, Bacon agrees with Boethius and Aquinas that words signify concepts. The latter, however, are not signified conventionally (and neither are they what words signify primarily). Rather, words signify concepts *naturally* (and things immediately):

> And it is clear for whoever asks, that after <a name> has been imposed exclusively upon an extra-mental thing [i.e. when the name was not imposed upon a concept], it is impossible that the vocal sound signifies the concept of <this> thing as a sign given by the soul and signifying conventionally, because a signifying vocal sound does not signify unless by imposition and institution. (Roger Bacon, *De signis*, § 163).
>
> When a thing is actually cognized and actually named by a speaker, the thing actually named and cognized implies [*ponit*] a concept [*species*] in the soul as well as a cognitive disposition, because a thing can only be cognized through its concept and if such a disposition exists in the soul; and if it cannot be cognized, it cannot be named significatively. Therefore, whenever a significative vocal sound is uttered significatively, a concept of a thing as well as a cognitive disposition must be actually present in the soul; therefore, a significative vocal sound uttered significatively and according to the convention [*ad placitum*] necessarily entails a concept of a thing as well as a cognitive disposition in the soul. But a natural sign in the first mode was understood in that way; therefore a vocal sound conventionally signifying a thing is a natural sign of the concept of the thing itself existing in the soul, and this according to the first mode of a natural sign. And Boethius explicitly says in his *Commentary on the book Perihermeneias* that a vocal sound signifies a concept of a thing, and Aristotle holds in the same place that vocal sounds are the marks of the impressions [*passiones*] which are in the soul, and such impressions are concepts [*species*] and dispositions according to Boethius, because a thing is not in the soul. (Roger Bacon, *De signis*, § 165) [22]

[20] Bacon's treatise is edited by Fredborg et al. (1978).

[21] On this specific point, see Maloney (1983); Meier-Oeser (1997, pp. 50–65); Rosier-Catach (1994, Chaps. 3–5); Marmo (2010, pp. 79–92 and 114–120).

[22] The first mode of a natural sign is that of the inference (a coloured sunset is a sign of rain the next morning); the two other modes of natural signs are similitude and causality (the effect is sign of its cause).

Bacon is of course perfectly aware of the fact that his interpretation twists the authoritative texts which lead quite naturally to the *mediantibus conceptibus* thesis as held, for example, by Aquinas. To the objection that this is not precisely what Aristotle and Boethius seem to have in mind he thus replies that there is a shift of perspective in the Aristotelian treatise: In the opening chapter, Aristotle is talking about signs in general—just as Bacon himself in most of the *De signis,* by the way...—and not specifically about linguistic signs. This is clearly shown, Bacon argues, by the fact that Aristotle presents concepts as natural signs of things (and concepts are certainly not linguistic signs); furthermore, it is only from Chap. 2 onwards—*De nomine*—that Aristotle does specifically speak about linguistic signs (and there, the signification of concepts by words is not a topic anymore).

As for the instrumental conception of words, already present in Abelard (and to a lesser extent in Aquinas), it plays a central role in Bacon although the idea is not expressed in the passages where he addresses the question of the relation between words, concepts, and things, but in the general characterization of the sign, given at the very beginning of his treatise on semiotics:

> The sign belongs to the category of relation, and it is said <to be a sign> essentially with respect to that [i.e. the intellect of a receiver] for which it signifies [*ad illud cui significat*], for it actually posits that <for which it signifies> [i.e. the intellect of the receiver] when it actually is a sign, and potentially, when it is a sign potentially. As a matter of fact, unless someone can conceive <something> by means of a sign [*concipere per signum*], it [i.e. the sign] would be vain and empty; moreover: it would not be a sign at all, but remain a sign only according to its substance and would not keep its essential function of sign [*ratio signi*], like a father's substance remains when <his> son is dead, but not the relation of paternity. (Roger Bacon, *De signis*, § 1)

Words, that is: linguistic signs, are used by speakers to make others conceive something by virtue of them. The essential role Bacon attributes to hearers (or more generally: to receivers, that is to "that *for which*" a sign signifies) is confirmed by many aspects of his theory of language,[23] and on one occasion at least, he even explicitly associates words with mechanical tools (de Libera and Rosier-Catach 1986, p. 63–79). Words are used by speakers to perform certain speech acts just like a stick is used to perform certain mechanical acts:

> much like a man or a soul is the main agent in the operation of negation and the word '*non*' the instrument, he who beats is the main agent in the act of beating and the stick is the instrument; and in the same way a man or a soul is the main agent in the <linguistic> subject's distribution [i.e. quantification], and '*omnis*' is the instrument. (Bacon 1937, pp. 153–154)[24]

[23] Contrary to the authors we considered so far, Bacon develops his views on language under a decisively Augustinian perspective (while the authority of Aristotle plays a comparatively minor role)—hence the overwhelming importance of the notion of communication and linguistic interaction in his semantics. On this point, see Rosier-Catach (1999).

[24] Such a concrete, instrumental conception of words can already be observed in the *Tractatus de proprietatibus sermonum* (ca. 1200, ed. Lambert Marie De Rijk, *Logica Modernorum*, II.2, p. 710): "But it seems that 'to signify' is not the same when it is said of a vocal sound or of a user <of language>.... One talks this way about beating [*percussio*], for the beating of the stick and the beating of the beating man are the same, but <the beating> of the stick <is> accidental as <is the

According to Bacon, then, concepts are of foremost importance in the semantics of words, but they are *not* what words signify (except in the very peculiar case of the imposition of words upon concepts). Concepts are that without which no thing—literally: *nothing*—could be understood, grasped, or conceived, and subsequently signified, or named. In other words, concepts are not *semantic mediators*, but sine qua non conditions for (a) the *potentially* successful use of language (on the part of speakers), and (b) the *actually* successful use of language (on the part of hearers).

26.3.3 Instead of Semantic Mediation: Semantic Subordination

William of Ockham (1287–1347), famous for his radical nominalism and the idea of a full-fledged mental language, quite fundamentally rethinks the role played by concepts *in significando* (on Ockham's philosophy of language, see Biard 1989, Chaps. 2–4, as well as Panaccio 1992, 1999). Like Bacon, he does reject the idea of concepts being primarily and conventionally signified by words, but defends a suprisingly different interpretation of Aristotle's words in the first chapter of the *De interpretatione:*

> But the Philosopher [i.e. Aristotle] says that a vocal sound *is a mark of an impression of the soul* [*passio animae*] because of a certain order prevailing among them in signifying; for a concept [*passio*] signifies things in the first place; in the second place, a vocal sound does not signify a concept, but the same things signified by the concept. (Ockham 1978, p. 347)

The parallel passage from the *Summa logicae* reads:

> I say that vocal sounds are signs subordinated to concepts or intentions of the soul [*intentiones animae*], not because those vocal sounds, taking the word 'signs' in its proper sense, always signify those concepts, but because vocal sounds are imposed in order to signify the same things which are signified by concepts of the mind [*conceptus mentis*], so that a concept in the first place naturally signifies something, and a vocal sound <conventionally> signifies the same thing in the second place. (Ockham 1974, pp. 7–8)

Aristotle and Boethius, claims Ockham, did not mean anything else when they said that vocal sounds signify concepts.[25] This reading radically differs from the traditional interpretation paradigmatically represented by Aquinas (Perler 1999). Concepts are not seen as intermediary steps in the semantics of words anymore, but words as well as concepts display parallel though subordinated semantics.[26] In

---

beating> of an instrument, whereas <the beating> of the user of an instrument <is said> in a proper way .... The same applies when one says: 'this one signifies a thing by means of a vocal sound' [*significat rem per vocem*], which means: 'he uses a sign and mark of a thing with the intention to produce a sign of the thing [*cum intentione faciendi signum de re*]". For a detailed discussion of this passage, see Rosier-Catach (1994, pp. 173–179).

[25] Ockham (1974, pp. 7–8): "Et universaliter omnes auctores, dicendo quod omes voces significant passiones vel sunt notae earum, non aliud intendunt nisi quod voces sunt signa secundario significantia illa quae per passiones animae primario importantur...".

[26] In that respect, Bacon seems to be quite close to the idea of subordination although he does not tackle it as such.

a certain sense, both Aquinas and Ockham would agree that words signify things *because* concepts signify things. But while for Aquinas the *explanans* refers to the necessary mediation of concepts—and thus to a meaning *process* or meaning *chain* characterizing the actual functioning of linguistic signs—for Ockham it expresses a *historical* and *epistemological* fact. On his account, a spoken word can conventionally signify things only if, at some point in the past, someone who had a concept naturally signifying these things on his or her mind decided to use the word at stake henceforth to signify precisely these very things.[27] Ockham's position also differs from that of Roger Bacon. Despite the fact that both authors argue for the non-mediated signification of things, their different conceptions of what a linguistic sign is make their approach radically different: For Bacon, the functioning of a spoken word as a sign is essentially dependent on a receiver and his or her mental activity (see the opening paragraph of the *De signis,* quoted above, § 26.3.2); for Ockham, by contrast, a spoken word is a sign essentially in virtue of its ability to stand for things (Ockham 1974, I.1; for a detailed discussion of Ockham's conception of the sign, see Panaccio 2004, Chap. 3).[28]

Ockham's originality also appears when one compares his positions with, for example, the ones of John Duns Scotus (1265–1308) and Walter Burley (1275–1344) who occupy, within the doctrinal development we are sketching here, an intermediate position between Aquinas and Ockham: Both thinkers hold that words signify things, but they insist that the *significata* of words are not things *simpliciter* (*res ut existunt*), but things insofar as they are cognized (*res ut intelliguntur* or *ut intellectae*); accordingly, the moment of mental mediation—by the *species* or mental images of things, a kind of entity forcefully rejected by Ockham (Tachau 1988, pp. 130–135)—remains essential in their semantic theories (on this question, see Pini 1999, 2001. See further Cesalli 2007, pp. 122–128 and 180–185).[29]

[27] Semantic subordination, as Claude Panaccio (2004, Chap. 9) has shown, occurs at the time of the imposition of names and does not require that the concepts to which spoken words are subordinated exist in the mind of a speaker at the time of the utterance (this is Ockham's so-called semantic externalism).

[28] Note, by the way, that Ockham's account of the semantic function of spoken words is remarkably close to what Karl Bühler, in opposition to Marty, will call "the coordination" (*Zuordnung*) of words with things and states of affairs: "There exists a totally different performance of language which cannot be derived from expressive moves and does not rest on the causal relation linking the uttered word to the hearer and the speaker, but depends on a relation which, in mathematics, is called coordination [*Zuordnung*]: the name is coordinated with its object, the statement with a state of affairs....This performance [*Leistung*] ... will be most adequately referred to...by the name of 'representation' [*Darstellung*]; for it is nothing but what images manage to perform with respect to certain states of affairs, or geographical maps with respect to others, namely: that the knower be able to grasp the state of affairs" (Bühler 1920, pp. 3–4).

[29] The position of Scotus seems to have changed over time, since in his *Ordinatio* (1304) he defends a position which is quite close to Ockham's (Panaccio 2004, p. 166).

26.3.4 A "Pragmatic Compromise": Semantic Mediation and Subordination

In the perspective of the present study, John Buridan (c. 1300–c. 1358) plays a notable role in that he combines the two hitherto competing ideas of the semantic mediation of concepts *and* of the semantic subordination of spoken to mental language (on Buridan's philosophy of language, see King 1985, pp. 1–84; Biard 1989, pp. 162–202; Zupko 2003, pp. 3–48; Klima 2009, Chaps. 2–9). In the preliminary remarks with which he opens the first treatise of his *Summulae,* Buridan makes the following observation:

> One has to know, therefore, that three discourses, three terms or <three> words can be distinguished to the extent that this point is touched upon at the beginning or the book *Peri hermeneias*, namely: mental, vocal and spoken <discourse>. ... One has to note as well that just like conventionally significative vocal sounds are related to mental concepts in signifying, written words are related to spoken words in signifying. Hence, spoken words do not signify extra-mental things unless through the mediation of the concepts to which they are subordinated [*voces non significant res extra nisi mediantibus conceptibus quibus subordinantur*], and neither do written words signify concepts or certain extra-mental things unless they signify the spoken words which refer to those concepts. (Buridan 2005, pp. 16–17)

However, whereas the conceptual mediation described here coincides with the traditional, pre-Ockhamian idea, subordination as Buridan understands it, is something quite different from what it means in Ockham. According to Buridan, the concepts of the things meant by speakers *must* occur in them *at the moment of utterance*—nothing can be signified, if it is not grasped conceptually. In other words, Buridanian semantic subordination is nothing but the backside of semantic mediation. This conception of signification goes hand in hand with the primacy of spoken over mental language. Like Roger Bacon, Buridan sees the essence of language in its communicative function:

> Regarding ... <signification>, one has to note that speech, or the power to utter vocal sounds [*virtus vociferandi*] was given to us in order to enable us to signify our concepts to others [*aliis significare conceptus nostros*], and the sense of hearing was given to us in order that the concepts of utterers be signified to us [*nobis significarentur conceptus vociferantium*]. ... Thus, it is evident that a signifying vocal sound has to signify a concept of the speaker to the hearer [*debet significare audienti conceptum proferentis*] and that it must evoke in the hearer a concept similar to <the one> of the speaker [*debet in audiente constituere conceptum similem conceptui proferentis*] ... <and> it is clear that the ones who discuss and speak intend precisely this, namely: that their vocal sounds operate [*operentur*] in those two ways. (Buridan 1998, p. 9)

Communication always occurs in a determinate situation and is highly context sensitive. Accordingly, the elucidation of linguistic meaning requires the taking into account of the complex circumstances of utterance, the pragmatic context, including the intentions of speakers (or writers):

> Signifying names are conventional; this is why different authors often use the same names equivocally, according to diverse intentions, and it is allowed for all authors who use names to interpret those names according to the intentions they have while using those names, and their hearers, as well as the readers of their books must receive the author's words accord-

> ing to the intention he has or seems to have. (John Buridan, *Summulae. De locis dialecticis*, VI.iii.3, text quoted in Biard 1989, p. 178)[30]

The "compromise" suggested by Buridan, then, can be qualified as "pragmatic" in a twofold sense: On the one hand, it combines the two traditional notions of semantic mediation and subordination ('pragmatic' here means something like 'ecumenical' or 'tactical'); on the other hand, it makes semantics depend on complex contextual factors by according a clear primacy to spoken over mental language and by ascribing a determinant role to speakers' intentions (here, 'pragmatic' is much closer to the contemporary, technical sense of the word).

26.4 Results

Which are the doctrinal similarities and differences between Anton Marty's understanding of the principle according to which *voces significant res mediantibus conceptibus* and the four medieval interpretative models presented above? In my opinion, this question must be answered in distinguishing two perspectives.

1. From a *strictly semantic* point of view, Marty's position does not correspond to any of the four medieval models, for these consider exclusively object-or entity-based semantics (words signify either concepts, or things, or the latter by the mediation of the former); by contrast, Marty's answer to the question of the nature of meaning (*Bedeutung/significatio*) in the proper sense of the expression is *not* given in terms of objects or entities (be they mental or extra-mental), but in terms of processes and norms. This, of course, also disqualifies Marty's naming (*Nennung*)—a semantic relation which is indeed mediated by *Bedeutung*—as a plausible equivalent of anything we found in our four medieval models.
2. From a *pragmatic–semantic* point of view, however, striking similarities appear between Marty and the medievals, for the idea that words function as tools used to express the speakers' concepts, and to evoke similar concepts in the hearers—in other words, the idea that *significare* also has the active sense of *intellectum constituere*—is present in three of the four medieval models we studied and it is, as we saw, at the very heart of Marty's semantics. The exception is constituted by Okham's conception of semantic subordination: just as Bühler with respect to Marty, Ockham with respect to the tradition before him (but also to Buridan) disconnects the function of concepts—more precisely, their role as that to which spoken words are subordinated—from the actual mental acts of speakers and hearers at the moment of utterance.

[30] On the same page, Biard quotes this eloquent passage from Buridan's *Quaestiones in Metaphysicen Aristotelis*, IX, 5 (ed. Paris, 1518, fol. 58va): "Sermones non habent virtutem nisi ex impositione et impositio non potest sciri nisi ex usu". Words do not possess any power unless by virtue of their imposition, while the imposition [i.e. the words' meaning] cannot be known unless through application [ex usu]—a claim which is in perfect accordance with Bacon's idea of the constant reimposition of words.

That said, the medieval model with which Marty's pragmatic semantics displays the strongest affinities is certainly that of Roger Bacon. For one thing, the account of language put forward by both authors is primarily based on the communicative interaction between speakers and hearers; secondly, they both accept the idea that concepts play a central role without claiming that concepts are what words signify. Furthermore, they both tightly link linguistic meaning with the actual mental acts of speakers and hearers. There is also a partial similarity between Bacon's interpretation of the first chapter of the *De interpretatione* and Marty's notion of indication (*Kundgabe*): Just as, according to Marty, the voluntary uttering of a name indicates (*gibt kund*) the existence in the speaker of a concept (*Vorstellung*) of the thing named; the uttering of a name, according to Bacon, is a natural sign of the presence of the concept of a thing in the mind of the speaker. However, Marty and Bacon disagree in that *Kundgabe* plays an essential role in the process of linguistic meaning, whereas the natural signification of the concept pointed at by Bacon is nothing but a side effect: The Baconian *significatio* of words does not depend on the fact that they naturally signify concepts.

Acknowledgments Many thanks to Nadja Germann for the careful reading of first version of this chapter.

References

Abelard P (1994) Morin P (ed/transl) Des intellections. Vrin, Paris
Abelard P (2010) Jacobi K, Strub C (eds) Glossae super Peri Hermeneias. Brepols, Turnhout
Ahlman E (1926) Das normative Moment im Bedeutungsbegriff. Druckerei der Finnischen Literatur Gesellschaft, Helsingfors
Amerini F (2000) La dottrina della significatio di Francesco da Prato O.P. (XIV secolo). Una critica tomista a Guiglelmo di Ockham. Documenti e Studi sulla Tradizione Filosofica Medievale 11:375–408
Aquinas T (1989) Expositio libri Peryhermeneias, I.2, Editio Leonina, t. I* 1, cura et studio fratrum praedicatorum. Vrin, Paris
Aristotle (1980) De interpretatione. English edition: Aristotle (1980), Aristotle's categories and propositions (De interpretatione) (transl: Apostle HG). Peripatetic Press, Grinnell Iowa
Bacon R (1937) Summa de sophismatibus et distinctionibus. In Steele R (ed) Opera hactenus inedita Rogeri Baconi, fasc. XIV. Clarendon, Oxford
Baumgartner W, Rollinger R, Fügmann D (eds) (2006/2009) Die Philosophie Anton Martys. Brentano Studien 12
Biard J (1989) Logique et théorie du signe au XIVe siècle. Vrin, Paris
Boethius (2010) Commanterius in librum Aristotelis Perihermeneias, editio secunda. English edition: Boethius (2010) On Aristotle On interpretation 1-3 (transl: Smith A). Duckworth, London
Bühler K (1920), Kritische Musterung der neuern Theorien des Satzes. Indoger Jahrb 6:1–20
Buridan J (1998) van der Lecq R (ed) Summulae. De suppositionibus, IV.i.2. Brepols, Turnhout
Buridan J (2005) van der Lecq R (ed) Summulae. De propositionibus, I.i.6. Brepols, Turnhout
Cesalli L (2007) Le réalisme propositionnel. Sémantique et ontologie des propositions chez Jean Duns Scot, Gauthier Burley, Richard Brinkley et Jean Wyclif. Vrin, Paris
Cesalli L (2013) Anton Marty's intentionalist theory of meaning. In Fisette D, Fréchette G (eds) Themes from Brentano. Rodopi, Amsterdam, pp 139–163

Chrudzimsky A (1999) Die Intentionalitätstheorie Anton Martys. Grazer Philos Stud 57:175–214
de Libera A, Rosier-Catach I (1986) Engendrement du discours et intention de signifier chez Roger Bacon. Hist Épistémol Lang 8(2):63–79
De Rijk LM (1967) Logica Modernorum, vol II.1. Van Gorcum, Assen
Fredborg KM, Nielsen L, Pinborg J (1978) An unedited part of Roger Bacon's Opus maius: 'De signis'. Traditio 34:75–136
Jacobi K, King P, Strub C (1996) From the intellectus verus/falsus to the dictum propositionis: the semantics of Peter Abelard and his circle. Vivarium 34(1):15–40
King P (1985) John Buridan's logic. Reidel, Dordrecht
Klima G (2009) John Buridan. Oxford University Press, New York
Magee J (1989) Boethius on signification and mind. Brill, Leiden
Maloney T (1983) Roger Bacon on the significatum of words. In: Brind'Amour L, Vance E (eds) Archéologie du signe. Pontifical Institute of Medieval Studies, Toronto, pp 187–212
Marenbon, J (2003) Boethius. Oxford University Press, New York
Marenbon J (ed) (2009) The Cambridge companion to Boethius. Cambridge University Press, New York
Marmo C (2010) La semiotica del XIII secolo. Bompani, Milano, pp 114–125
Marty A (1884/1018) Über subjectlose Sätze. In: Eisenmeier J, Kastil A, Kraus O (eds) Anton Marty. Gesammelte Schriften, Bd II.1. Max Niemeyer, Halle
Marty A (1893/1920) Über das Verhältnis von Grammatik und Logik. In: Eisenmeier J, Kastil A, Kraus O (eds) Anton Marty. Gesammelte Schriften, Bd II.2, Max Niemeyer, Halle, pp 57–99
Marty A (1894/1918) Über subjectlose Sätze und das Verhältnis der Grammatik zu Logik und Psychologie. In: Eisenmeier J, Kastil A, Kraus O (eds) Anton Marty. Gesammelte Schriften, Bd II.1. Max Niemeyer, Halle
Marty A (1908) Untersuchungen zur Grundlegung der allgemeinen Grammatik und Sprachphilosophie. Max Niemeyer, Halle
Meier-Oeser S (1997) Die Spur des Zeichens. Das Zeichen und seine Funktion in der Philosophie des Mittelalters und der frühen Neuzeit. De Gruyter, Berlin
Mulligan K (ed) (1990) Mind, meaning and metaphysics: the philosophy and theory of language of Anton Marty. Kluwer, Dordrecht
Noiré L (1882) Die Lehre Kants und der Ursprung der Vernunft. J. Diemer, Mainz
Ockham W (1974) Boehner P, Gál G (eds) Summa logicae, St. Bonaventure University, St. Bonaventure
Ockham W (1978) Gambatese A, Brown S (eds) Expositio in librum Perihermeneias Aristotelis, St. Bonaventure University, St. Bonaventure
Panaccio C (1992) Les mots, les concepts et les choses. La sémantique de Guillaume d'Occam et le nominalisme contemporain. Bellarmin, Montréal
Panaccio C (1999) Le discours intérieur. De Platon Ã Guillaume d'Ockham. Le Seuil, Paris
Panaccio C (2004) Ockham on concepts. Ashgate, Aldershot
Perler D (1999) Direkte und indirekte Bezeichnung. Die metaphysischen Hintergründe einer semantischen Debatte im Mittelalter. Bochumer Philosophisches Jahrbuch für Antike und Mittelalter 4:125–152
Pinborg J (1971) Bezeichnung in der Logik des 13. Jahrhunderts. In: Zimmermann A (ed) Der Begriff der repraesentatio im Mittelalter. De Gruyter, Berlin, pp 238–281
Pini G (1999) Species, concept, and thing: theories of signification in the second half of the thirteenth century. Mediev Philos Theol 8:21–52
Pini G (2001) Signification of names in Duns Scotus and some of his contemporaries. Vivarium 39(1):21–51
Rollinger R (2010) Philosophy of language and other matters in the work of Anton Marty. Rodopi, Amsterdam
Rosier-Catach I (1994) La parole comme acte. Sur la grammaire et la sémantique au XIIIe siècle. Vrin, Paris
Rosier-Catach I (1995) Henri de Gand, le De Dialectica d'Augustin, et l'imposition des noms divins. Documenti e studi sulla tradizione filosofica medieval 6:145–253

Rosier-Catach I (1999) Aristotle and Augustine. Two models of occidental medieval semantics. In: Sing Gill H, Manetti G (eds) Signification in language and culture, vol II. Bahri Publications, New Dehli, pp 41–62
Tachau T (1988) Vision and certitude in the age of Ockham. Optics, epistemology, and the foundation of semantics 1250–1345. Brill, Leiden
Valente L (2008) Logique et théologie dans les Écoles parisiennes entre 1150 et 1220. Vrin, Paris, pp 36–62
Zupko J (2003) John Buridan. University of Notre Dame Press, Notre Dame

Chapter 27
Mental Files and Identity

François Recanati

Abstract Mental files serve as individual or singular concepts. Like singular terms in the language, they refer or are supposed to refer. What they refer to is not determined by properties which the subject takes the referent to have (i.e. by the information stored in the file), but through relations to various entities in the environment in which the file fulfils its function. Files are based on *acquaintance relations,* and the function of the file is to store whatever information is made available through the relations in question.

I offer a typology of files. The most important distinction is between proto-files and conceptual files. In contrast to proto-files, conceptual files can host not only information derived through the specific relation on which the file is based but also information about the same object gained in some other way.

In this framework, identity comes into the picture twice: (a) Identity is presupposed when two pieces of information occur in the same file. Such 'presumptions of identity' ground the linguistic phenomenon of *de jure* coreference, which takes place when two singular terms, or two occurrences of a singular term, are associated with the same file. (b) Judgments of identity work by linking two *distinct* files, thereby enabling information to flow freely between them. This corresponds to *de facto* coreference. (Linking is not merging; identity judgments have the effect of merging files only when the files belong to a very specific category, that of 'encyclopedia entries'—a type of conceptual file based on a higher-order relation rather than on a specific acquaintance relation.)

In the last part of the chapter, I discuss, and attempt to rebut, two objections to the mental-file account. According to the first objection, the account is circular; according to the second objection, *de jure* coreference cannot be accounted for in terms of identity of the associated mental files because *de jure* coreference is not a transitive relation.

Keywords Modes of presentation · Mental files · Identity · Coreference *de jure* and *de facto* · Singular thought

F. Recanati (✉)
Institut Jean-Nicod, CNRS-ENS-EHESS, Paris, France
e-mail: francois.recanati@ehess.fr

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_27,

27.1 Introduction

A rational subject, *S,* may take different (and possibly conflicting) attitudes towards the judgment that a given individual is *F*—for example, she may reject it as false or accept it as true—depending on how that individual is presented. For one and the same individual *x,* say Cicero, *S* may accept the claim that *x* was a philosopher if that claim is made in a certain way ('Cicero was a philosopher'), while rejecting the claim that *x* was a philosopher if it is made in a different way ('Tully was a philosopher'). Both 'Cicero was a philosopher' and 'Tully was a philosopher' say of the individual to whom both 'Cicero' and 'Tully' refer that he was a philosopher, so they make the same claim (true iff the individual in question was a philosopher), but the subject's acceptance or rejection of the claim depends upon the mode of presentation of the referent the two names share. If the referent is presented as Cicero, the claim is accepted, but if he is presented as Tully, it is not. That, of course, is possible only if *S* does not realize that Cicero *is* Tully. I assume that *S* has both the names 'Cicero' and 'Tully' in her repertoire, and that both names, as she uses them, refer to one and the same individual. The problem is that *S* herself does not know that. For her, there are two distinct individuals and two distinct claims are made (one with respect to each of them).

To account for that sort of situation, Frege posited modes of presentation, or 'senses', in addition to the reference of linguistic expressions. And he appealed to this idea to account for the informativeness of identity statements such as 'Cicero is Tully'. At the level of reference, the statement is trivial, since an individual (the common referent of 'Cicero' and 'Tully') is said to be identical to that very individual—hardly a contingent matter. At the level of sense, however, the statement is informative precisely because the senses associated with 'Cicero' and 'Tully' are distinct. Sense determines reference but does so only contingently. Because of that element of contingency, it is not guaranteed that the referents determined by two distinct senses (e.g. the sense of 'Cicero' and the sense of 'Tully') will be the same, and indeed a subject like *S,* unaware of certain contingent facts, takes them not to be the same. If the *senses* of the names were themselves the same (as in 'Cicero is Cicero'), the statement would be trivial and recognized as such by whoever understands it.

Now what *are* senses or modes of presentation? Frege himself thought of them as essentially descriptive. The referent is presented as having certain properties or standing in certain relations to other entities. Since sense is supposed to determine reference, a unique object must have the relevant properties or stand in the relevant relations to other entities. So a sense can, in principle, be expressed by means of a *definite description* 'the F'. The unique object which satisfies the descriptive condition ('F') is the referent. In cases such as 'Cicero' and 'Tully', the suggestion is that the subject, *S,* associates different descriptions with the two names.

There are good and well-known reasons for rejecting Frege's descriptivist construal of senses, however. Senses are supposed to help us account for 'Frege cases'. In Frege cases (e.g. 'Hesperus'/'Phosphorus', 'Cicero'/'Tully', etc.), the cognitive

significance of two terms differs even though their reference is the same. (As a result, a rational subject may be led to ascribe contradictory properties to what is in fact the same object.) The problem with Frege's descriptive take on senses is that, if accepted, it forces the theorist to posit reference-determining descriptions in the head of the subject *whenever* a Frege case is possible. Now there are three types of case in which contemporary philosophers of language and mind have been reluctant to do so:

1. *Reference through acquaintance*. When we perceive an object and have a thought about it, the object the thought is about is the object the perception is about; and that, arguably, is not determined by properties the subject takes the referent to have (Pylyshyn 2007). In many cases, we are actually unable to properly describe the object that is given to us in experience: We do not know what it is yet that does not prevent us from referring to it directly (without conceptual mediation) and wondering what *it* can be (Dretske 1988, p. 73).[1] In such cases, even though we are unable to conceptually articulate what our thought is about, Frege cases are still possible. I may be perceptually related (through distinct sense modalities, say) to what I take to be two objects, which happen to be one and the same object.

Faced with such cases, the Fregean is likely to say that the reference-fixing description in the mind of the subject must be something like 'what I am now seeing' or 'what I am now touching'. But this supposes, on the part of the subject, reflective abilities the exercise of which is not required to suffer from identity confusions of the type which Frege cases illustrate. The subject need not reflect on her perceptual relation to objects in order to have thoughts about the objects she perceives, nor does she have to reflect on her perceptual relation to objects to be in a position to think of the object in different ways, corresponding to the various ways in which she perceives it.

2. *Reference through communicative chains*. In cases such as the 'Cicero'/'Tully' case, the subject is able to describe the referent but the descriptions he or she can provide do not fill the Fregean bill. First, the descriptions the subject can provide are often indefinite ('a famous Roman orator') rather than definite. That does not prevent the term(s) from referring. Second, when the subject is able to provide a definite description, the descriptive condition often fails to be satisfied by a unique object. Again, that does not prevent the term with which the description is associated from referring. Third, assuming the description the subject can provide is definite and uniquely satisfied, the satisfier need not be the referent of the term whose sense we are trying to characterize, as in Kripke's Gödel/Schmidt case. This type of consideration has led a number of theorists, in the late sixties and early seventies, to argue in favour of an 'externalist' approach to reference determination. According to Geach (1972), Kripke (1980), Donnellan (1970)

[1] As Campbell (2006, p. 205) puts it, 'Your visual system is managing to bind together information from a single thing, and you are consequently able to attend consciously to it, even though you have not managed to apply the right sortal concept to it'.

and others, what determines the reference of a name on a given use is not a description in the head of the users but historical facts about that use and the communicative chain to which it belongs.

At this point, the Fregean can make the following response: The description in the mind of the users in that sort of case is something like 'the person called *Cicero*', a description which (if the historical-chain picture is correct) is satisfied by whoever stands at the other end of the communicative chain which eventuates in the current use of the name. As Peter Geach pointed out, however, the *existence* of a communicative chain is sufficient to enable a name user to successfully refer. The communicative chain does not have to be represented, any more than the perceptual relation to the referent has to be represented, in order for the subject to successfully refer to an object he is acquainted with.

3. *Refer*ence *through indexicals*. Indexical modes of presentation are essentially perspectival and cannot be captured by means of objective, non-indexical descriptions. As Castañeda and (following him) Perry forcefully pointed out, for any indexical *a*' and non-indexical description 'the F', it is always possible for the subject to doubt, or to wonder, whether *a* is the *F* (Castañeda 1999; Perry 2000).

Reichenbach has suggested that an indexical is equivalent to a token-reflexive description (Reichenbach 1947, § 50). Thus, a given token of 'I' presents its referent as the utterer of that token, a token of 'now' presents the time it refers to as including (or overlapping with) the time at which this token is uttered, etc. Insightful though it is, this move cannot support a descriptivist approach to indexical modes of presentation. What is needed to support such an approach is an objective, *non-indexical* description that provides the sense of the indexical. But for the token-reflexive description to count as non-indexical, the token in terms of which the referent is described must itself be described in objective/non-indexical terms, rather than referred to by means of a demonstrative like 'this token' (itself a variety of indexical). Now if the token is objectively described as, say, 'the F-token', the token-reflexive description will no longer be suitable for capturing the sense of the indexical. It is certainly possible for me to doubt that I am uttering the F-token, or to doubt that the F-token is being uttered now (or here), and that is sufficient to establish that the token-reflexive description ('the utterer of the F-token', 'the time/place at which the F-token is uttered'…) does not provide the sense of the corresponding indexical ('I', 'here' or 'now'). In any case, such token-reflexive descriptions can only be grasped by fairly sophisticated users of the language who, in addition to mastering the notion of 'token', are able to reflect upon the relations between token representations and objects in the context in which these representations occur. Indexical thinking indeed exploits these relations but in no way presupposes the ability to reflect on them.

To handle these three types of cases, which are counterexamples to Frege's descriptivist approach, we should make room for *nondescriptive* senses or modes of presentation. But what are such modes of presentation? What is it to think of an object nondescriptively? The theory of mental files is meant to answer that question. A nondescriptive mode of presentation is said to be a mental file. In Frege cases,

the subject has two mental files for the same object (without realizing that there is a single object).

27.2 Mental Files

27.2.1 Mental Files and Epistemically Rewarding Relations

There are different types of mental file, as we shall see in due course (Sect. 27.3) but they all have the following properties. Mental files serve as individual or singular concepts. Like singular terms in the language, they refer, or are supposed to refer. What they refer to is not determined by properties which the subject takes the referent to have, but in externalist fashion, through relations (of the subject, or of the file itself construed as a mental particular) to various entities in the environment in which the file fulfils its function. The primary function of the file is to store information from these entities—information that is made available through the relations in question. The information (or misinformation) in the file, therefore, corresponds to the properties which the subject takes the referent to have, but these properties are not what determines the reference of the file.

The characteristic feature of the relations on which mental files are based, and which determine their reference, is that they are *epistemically rewarding* (hence my name for them: ER relations). They enable the subject to gain information from the objects to which he stands in these relations. In all the cases mentioned above as objecting to Frege's descriptivist approach, ER relations are involved. Relations of perceptual acquaintance are ER relations: They are the sort of relation to objects which makes the perceptual flow of infomation possible. Thus, by holding an object in my hand, I can get information about its weight; by looking at it I can get information about its visual appearance. Perceptual files are, to use Perry's analogy, 'buffers' in which we store the information gained on the basis of these short-term relations. The relations of 'mediated acquaintance' established through communicative chains are also ER relations, which enable the subject (through communication) to gain information from the object at the other end of the communicative chain. The corresponding files are more enduring than perceptual buffers because the ER relation established through a communicative chain lasts longer than a transient perceptual relation.

The contextual relations to objects which indexical reference exploits are also ER relations, and they are typically short lived, but not always. According to Perry (2002), the SELF file (which provides the sense of the indexical 'I') is based upon a special relation which every individual (permanently) bears to himself or herself, namely identity. In virtue of *being* a certain individual, I am in a position to gain information concerning that individual in all sorts of ways in which I can gain information about no one else, e.g. through proprioception and kinaesthesis. The mental file SELF serves as a repository for information gained in this way. In contrast, the files associated with the other indexicals ('here', 'now'…) are based on short-lived

ER relations to the place we are in, or to the current time, which relations enable the subject to know (by using his senses) what is going on at the place or time in question. They are similar to perceptual buffers, which is to be expected given the link between indexicality and perception.

On the mental-file picture, what distinguishes descriptive from nondescriptive senses is the mechanism of reference determination. To use Kent Bach's useful terminology, reference determination is 'satisfactional' in the descriptive case and 'relational' in the nondescriptive or *de re* case:

> Since the object of a descriptive thought is determined satisfactionally, the fact that the thought is of that object does not require any connection between thought and object. However, the object of a *de re* thought is determined relationally. For something to be the object of a *de re* thought, it must stand in a certain kind of relation to that very thought. (Bach 1987, p. 12; see also Bach 1986, pp. 188–189)

There is a variety of descriptivism which accommodates the relational nature of *de re* thought but 'internalizes' the relations by incorporating them into the content of the associated descriptions. On this view, the sense of a singular term always is that of a definite description but in the allegedly 'nondescriptive' cases the descriptive condition *F* is relational. Thus, as we have seen, the descriptivist can say that in the perceptual case the mode of presentation is something like 'what I am seeing' or 'what I am touching'[2]; and similarly for the other cases that raise prima facie difficulties for the descriptivist. But, as we have also seen, this view supposes reflective abilities the exercise of which is not actually required for having the relevant thoughts. The mental-file picture avoids this intellectualist pitfall. Mental files are based on relations to objects. Their function is to store information gained in virtue of standing in that relation to objects, and to represent them in thought. By deploying the file (or its 'address' or 'label') in thought, the subject can think about the object in virtue of standing in the relevant relation to it. But to entertain the thought, the subject does not have to reflect upon the relation in which she stands to the object.

27.2.2 Linking Files

Identity statements are informative, Frege says, whenever the senses of the terms on each side of the identity sign are distinct. Thus, 'A=B' is informative, in virtue of the distinctness of the relevant senses, while 'A=A' is not, since (presumably) the same sense is exercised twice. In mental-file talk, this translates as follows: An identity statement 'A=B' is informative to the extent that the terms 'A' and 'B' are associated with distinct mental files. If the two terms are associated with the same file, the statement reduces to a (trivial) assertion of self-identity.

To say that there are two distinct mental files is to say that information in one file is insulated from information in the other file. Files are a matter of information

[2] Or, in token-reflexive form: 'what is causing this visual/tactile experience' (see, e.g. Searle 1983 for a token-reflexive analysis of the content of perceptual experience).

clustering. Clustering takes place when all the information derives from the same source, through the same ER relation, and when it takes place it licences the integration and inferential exploitation of the information in question. The role of the file is precisely to treat all the information as if it concerned one and the same object, from which it derives.[3] But integration and exploitation of information is blocked if the relevant information is distributed in distinct files, for then, there is no presumption that all the information derives from the same object. So, even if I know that Cicero is bald, and that Tully is well read, I cannot conclude that some bald man is well read, despite the fact that Cicero is Tully: The information 'is bald' is in the Cicero file, while the information 'is well read' is in the Tully file. Informational integration and inferential exploitation of information only takes place within files, on this picture.

There is, however, an operation on files whose role is precisely to overcome that architectural limitation, by licensing the integration/exploitation of information distributed in distinct files. That operation, following Perry, I call *linking*. When two files are linked, information can flow freely from one file to the other, so informational integration/exploitation becomes possible (despite the constraint that it can only occur within files). Thus, if I learn that Cicero is Tully, this allows me to put together the pieces of information in the two files, and to infer that some bald man is well read.

From a cognitive point of view, linking is a quite fundamental operation. It is involved, for example, in the phenomenon of recognition (which involves linking a perceptual file and a file based on memory). It is that operation which I think accounts for the cognitive effect of accepting an identity statement. To accept the identity 'A = B' *is* to link the two files corresponding to the terms on each side of the equal sign. It would be incoherent to accept the identity 'Cicero = Tully', and not let the information in the respective files get together and breed.

27.2.3 Linking Versus Merging

Strawson was the first to claim that identity judgments should be understood in terms of their effects on the management of files in the mind of the thinker (Strawson 1974, pp. 51–56). Two 'segregated bundles or clusters of identifying knowledge', he says, are 'brought together and tied up into one for a given audience of an identity statement' (52). But the operation on files which, according to Strawson, results

[3] 'Updating one's files involves being disposed to collect information as if there is some one individual that one's file F has always been about. One's screening and pruning dispositions are responsive to this purported fact' (Lawlor 2001, p. 88). 'One reason for not allowing an individual concept [=a file] to change its referent is that the referent fixes a condition for the coherence of information within an individual concept: if "is F" belongs in belief mode to a given individual concept, then "is not F" should not. The constraints on updating would not obtain if an individual concept might shift its referent' (Sainsbury 2005, p. 232). For a detailed argument to the same effect, see Millikan (1997, pp. 504–506; 2000, pp. 141–144).

from accepting an identity statement is the 'merge' operation through which the two files become one. ('Merge' is Millikan's term, not Strawson's). As several authors noticed, however, the 'merge' model is not adequate to describe the cognitive effects of identity judgments.

Two linked files may end up being merged, after some time (especially as new information accumulates), but there are all sorts of reasons also for not automatically merging two files that are linked (Lawlor 2001, pp. 62–65 and 92–93). For example, it would be very risky to merge two files on the basis of an identity judgment that one may accept with less than 100 % subjective probability (Millikan 1997, p. 508). Linking is less risky, as it can easily be undone. So merge is an option for dealing with an identity but it should not be automatic.[4] Second and most importantly, the 'merge' model is incompatible with the mode of presentation idea we are trying to cash out (Millikan 2000, pp. 147–149). It is of the essence of modes of presentation that there can be a multiplicity of modes of presentation for the same object. On the picture I have presented, mental files qua nondescriptive modes of presentation correspond to various relations in which the subject stands to objects, and there is no doubt whatsoever that a subject can and typically does stand in several relations simultaneously to the objects in his or her environment. Nor is this situation contrary to some normative ideal, as if the coexistence of several files for a single object was a defect to be avoided whenever possible. Imagine that I see a certain man cutting his grass and recognize him as Noam Chomsky. (Or imagine I learn he is Noam Chomsky, through an identity statement which I accept.) My perceptual file and my Chomsky file get linked, but there is no reason why either should disappear. Perry describes the perceptual buffer as being 'absorbed' into the more permanent file in such cases, but I think the buffer should only disappear when the ER relation on which it is based no longer holds. Taking seriously the idea that mental files are modes of presentation based upon contextual relations to objects demands that we accept the existence of a multiplicity of files for the same object, even when the files are linked and the subject is aware that they stand for a single object.

Strawson's idea that two linked files should be merged makes sense in the context of his own enquiry, however. He was concerned with a very specific type of file, associated with proper names (his topic in the relevant passage of *Subject and Predicate in Logic and Grammar*). I call that type of file an 'encyclopedia entry' (Recanati 1993, 2010; § 3.3 below). Encyclopedia entries do obey the norm that there should be exactly one per object of interest. In *Reference and Reflexivity,* Perry describes files as 'little cards in the mind on which we jot down information about people, things and places. My picture is *a card for each person, place, or thing*' (Perry 2001b, p. 54; emphasis mine). This description fits encyclopedia entries well, though I think it is a mistake to apply it to files in general. I shall say more about encyclopedia entries, and the norm that governs them, in Sect. 3. For my present purposes, the important point is that encyclopedia entries are only one particular type of file, and that the norm in question only applies to that type of file. It

[4] 'If the identification [A = B] is tentative, the notions [= files] may retain their identity; if not, they may merge and become one' (Perry 2002, p. 196).

does not apply to files in general, hence there is no reason to accept that, in general, linking does or should give rise to a merging of files.

27.2.4 Presumptions of Identity Within Files

Besides judgments of identity, the cognitive effect of which is to link two files, there are also *presumptions* of identity, whose status is quite different. While linking only operates on distinct files, presumptions of identity are operative within a single file. As I wrote above, to put various pieces of information in the same file means that they are suppose to derive from a single source (i.e. through some ER relation to a given object). Pieces of information in the same file can thus be inferentially integrated and exploited as if they concern the same object (whether or not they actually do). This Campbell describes as 'trading on identity' (Campbell 1987, 1994, 2002; see also Millikan 1997; Fine 2007).

It is tempting to regard presumptions of identity as nothing but *implicit* judgments of identity. On this view the difference between argument *A* and argument *B* below is that, in argument *A,* the judgment of identity is explicit while it remains implicit in argument *B* (which is therefore enthymematic). The reason why it can remain implicit in *B* is that the identity is obvious and trivial, so it 'goes without saying' and can be suppressed, in contrast to what happens in argument *A*.

Argument A	*Argument B*
Cicero is bald	Cicero is bald
Tully is well-read	Cicero is well-read
Cicero = Tully	[implicit premise : Cicero = Cicero]
---	---
Someone is bald and well-read	Someone is bald and well-read

The suppressed premise in argument *B* is meant to ensure that the two occurrences of 'Cicero' in the explicit premises of the argument actually corefer (if they did not corefer, the argument would be invalid, indeed). As Campbell and many others have shown, however, this view of argument *B* as enthymematic and resting on a suppressed premise is indefensible. In general, the attempt to reduce presumptions of identity to implicit identity judgments launches an infinite regress:

> If this view were correct, we would also need to make sure that the uses of ['Cicero'] in the suppressed premise are linked with the uses of ['Cicero'] in the explicit premises, and we would need further suppressed premises to secure these connections. The problem recurs, and we are embarked on a regress. (Campbell 1994, p. 75)
>
> This means that we cannot regard arguments like the one under consideration as enthymematic, needing but a further (object-language) sentence to be made completely valid; there is no evading unthinking reliance on sameness of reference. (Sainsbury 2002, p. 135)

> According to the [suggestion]… what it is to think that the individual Cicero is a Roman and then to have the coordinated thought that he is an orator is to think the additional thought that the one individual is the same as the other. But if the new thought is to have the desired effect, then it must be supposed that the individuals in the new thought are represented as the same as the respective individuals in the original thoughts; and so the account is circular. (Fine 2007, p. 68)
>
> To insist that the subject must make an explicit identity judgment before she can recognize that two thoughts are about the same thing would be to invite a vicious regress—for even the simplest inference from 'P' to 'P' would then require infinitely many explicit identity judgments to establish the co-reference of premise and conclusion. The moral here is much the same as the one Lewis Carroll drew in the case of modus ponens: we must have some basic way of taking two thoughts to be co-referential which does not require an explicit identity judgment. (Schroeter 2008, p. 115n)

I conclude that we need both the identity presumptions (or, according to the now standard terminology, 'coreference *de jure*') and identity judgments ('coreference *de facto*').[5] In the mental-file framework, they correspond to the clustering of information into files, and the linking of (distinct) files.

27.3 The Hierarchy of Files

27.3.1 Proto-Files and the Generality Constraint

I have alluded to the existence of several types of file. So far, the files I have talked about (with the exception of 'encyclopedia entries', on which more below) are very closely tied to specific ER relations on which they are based. The file exists only as long as the relation (hence the possibility of gaining information about the object by exploiting the relation) exists, and for that reason the life expectancy of many files is rather short: They are *temporary* files. So, as long as I am in the right type of perceptual contact with the grass-cutting man, I can think of him demonstratively ('that man'). When I am no longer in a position to perceive him or to focus my attention on him, I can no longer think of him under the demonstrative mode of presentation, since the latter involves the activation of a mental file which depends upon the existence of the right type of perceptual relation. When the relation is broken, the temporary file based on it disappears. (The information in the file is not

[5] Fine (2007) distinguishes 'thinking of something as the same' (*de jure* coreference, in the standard terminology, or, in Fine's own terminology, 'strict coreference') and 'thinking of something as being the same'. 'A good test of when an object is represented as the same is in terms of whether one might sensibly raise the question of whether it is the same. An object is represented as the same in a piece of discourse only if no one who understands the discourse can sensibly raise the question of whether it is the same. Suppose that you say "Cicero is an orator" and later say "Cicero was honest," intending to make the very same use of the name "Cicero." Then anyone who raises the question of whether the reference was the same would thereby betray his lack of understanding of what you meant' (Fine 2007, p. 40). However, as we shall see in Sect. 27.4, *de jure* coreference as characterized through this test is an overly broad notion which covers different types of case, including some cases in which distinct files are involved.

lost, of course, but transferred into other files, e.g. the Chomsky file to which the demonstrative buffer is linked.)

At this point, we need to introduce a distinction between two types of files based on ER relations. 'Proto-files', as I am going to call them, can *only* host information gained in virtue of the ER relation to the referent. For example, the proto-file SELF* can only host information gained 'from inside', in the first-person way; the demonstrative proto-file THAT MAN* can only host information gained by perceptually attending to the object. I call these files 'proto-files' (and mark them with an asterisk) rather than files *simpliciter* because I want files (properly speaking) to serve as individual *concepts,* i.e. thought constituents; and I take proto-files to lack a distinguishing characteristic of concepts.

Concepts, in general, satisfy or ought to satisfy what Evans calls the Generality Constraint. Evans says that a subject in possession of a predicative concept *F* should be able to entertain thoughts in which that concept is applied to any object of which the subject has an individual concept; similarly:

> If a subject can be credited with the thought that *a* is *F*, then he must have the conceptual resources for entertaining the thought that *a* is *G*, for every property of being *G* of which he has a conception. This is the condition that I call 'The Generality Constraint'. (Evans 1982, p. 104)

In the mental-file framework, predication is cashed out as follows: The file is what stands for the object of which something is predicated, and the predicate's location within the file means that it is taken to apply the object in question. Translated into mental-file talk, the Generality Constraint says that a file should be hospitable to any predicative concept in the subject's possession. Clearly, that is a constraint which proto-files do not satisfy. Take the proto-file SELF*: It can only host information gained from inside, through e.g. proprioception or introspection. Now, there is much information about myself that I cannot gain in this way. My date of birth is something I learn through communication, in the same way in which I learn my parents' birthdates. In virtue of the Generality Constraint, it should be possible for that information to go into my self file, and that is the crucial difference between the self file and the (nonconceptual) proto-file SELF* from which it originates.

27.3.2 Conceptual Files

In contrast to proto-files, which are based on some ER relation and can only host information derived through that relation, a (conceptual) file based on a certain ER relation contains two sorts of information: information gained in the special way that goes with that relation (first-person information, in the case of the SELF file), and information not gained in this way but *concerning the same individual as information gained in that way*. Information about my birthdate is a case in point: I gain that information in a third-person way, through communication (as I might come to know anybody's birthdate), but I take that piece of information to concern the same person about whom I also have direct first-person information, i.e. myself;

so that information, too, goes into the self file. I am therefore able to exercice my self-concept in thinking 'I was born in 1952'.

It is because of that dual aspect of the self-concept *qua* satisfier of the Generality Constraint that there are two types of 'I'-thoughts: those that are, and those that are not, immune to error through misidentification. When some information is gained from inside, that is, in virtue of the ER relation on which the SELF file is based, that information can only be about the subject: the way the information is gained determines which object it concerns (or, equivalently, in which file it goes). As a result, as Evans puts it, 'there just does not appear to be a gap between the subject's having information (or appearing to have information), in the appropriate way, that the property of being *F* is instantiated, and his having information (or appearing to have information) that *he* is *F*' (Evans 1982, p. 221). But when some information about ourselves is gained from outside, it goes into the SELF file only in virtue of a judgment of identity. The thought 'I was born in 1952' can thus be seen as the product of two thoughts: The thought that a certain person *x,* namely the person I hear about in a given episode of communication, was born in 1952, and the thought that I am *x*. The thought that I was born in 1952 thus turns out to be 'identification-dependent', in Evans' terminology.

Despite the fact that they can host information not derived through the ER relation on which the file is based, the file is still based on that relation. What this means is that the file exists only as long as the relation exists (and with it the special way of gaining information about the referent through the relation). Files should, therefore, be seen as an expansion of proto-files, including the proto-files themselves as their nucleus.[6] Linking is what necessitates the expansion of the original proto-files. Linking makes information flow between files, but that only makes sense if the files can accept 'alien information' (information from other files, that is, information that is not derived through the ER relation on which the file is based). So linking and expansion of proto-files are best construed as two sides of the same coin.

27.3.3 Encyclopedia Entries

The distinction between proto-files and conceptual files based on the Generality Constraint was drawn in *Direct Reference* (Recanati 1993, Chap. 7). Another important distinction, also made in Chap. 7 of *Direct Reference,* is between two types of conceptual files. The files I have talked about so far are what we might call 'first-order' files. They are based on specific ER relations. But we must allow room for more abstract files, based on the relation to an object that holds when we have first-order files about it, that is, when we stand in various ER relations to it which enable

[6] Peacocke notes that when the subject falls prey to a perceptual illusion of which he is aware, the content of the illusion ought not to figure in the subject's conceptual file about himself, because, at the level of judgment, 'the subject rejects the content of his more primitive, pre-judgemental phenomenology' (2012: 84). This might be taken to argue against the inclusion of the primitive proto-file within the conceptual file. I acknowledge the difficulty, but cannot discuss the issue here.

us to gain information from it in various ways. In Recanati (2010) I characterize such files as follows:

> Not all files are based on specific contextual relations enabling us to gain information about the referent in particular ways. Some files (the *indexical* files) are based on specific contextual relations, such as one's relation of identity to oneself or the relation to what we hold in our hand, but others (the *encyclopedic* files) are based on a more general-purpose tracking relation. Thus my file about Mont Blanc contains all the information I can get about the mountain, *however it is gained*. It is not tied to a particular way of gaining information, nor to a specific ER relation. An encyclopedic file may exploit a number of ER relations to the reference of the file, in an opportunistic manner, instead of being based on a single one. Any relation will do, provided it preserves the link to the object. In this case, what determines the reference of the file is the overarching tracking relation: the relation between the file and the object it has been created to track (however it *is* tracked). Not being based on a specific ER relation, an encyclopedia entry is not short-lived, as the other type of file typically is. It survives when our contextual relation to the reference changes. (Recanati 2010, pp. 157–158)

This might suggest that encyclopedic files are not based on ER relations; however, that is not what I want to say. Rather, I distinguish between specific ER relations and the higher-order ER relation on which encyclopedia entries are based, namely

$$\lambda x \lambda y\ [(\exists R)\ (Rx, y)]$$

where 'R' ranges over ER relations. A subject (or a mental file in the subject's mind) x stands in that relation to an object y just in case there is/are some first-order ER relation(s) in which x stands to y. A file based on the higher-order relation hosts any information derived in virtue of that relation, that is, ultimately, any information derived in virtue of any of the first-order ER relations. Such files correspond to what Perry calls 'detached' files.[7]

For encyclopedia entries, the Strawsonian constraint 'one object, one file' holds: Since the file abstracts from the specific ER relations, there is no point in entertaining distinct encyclopedia entries about the same object. All such files would be based on the same relation to the object (the higher-order relation) so their multiplicity could only reflect the mistake of thinking that there are two objects where there is one.

I speak of a 'hierarchy' of files for two reasons. First, files at each of the three levels I have described presuppose files at the previous level. Proto-files are the most basic; conceptual files are generated from them (through linking-*cum*-expansion). Among conceptual files, first-order files are more basic, since higher-order files—encyclopedia entries—presuppose them. Second, files can be ordered in terms of how closely tied they are to specific ER relations. Proto-files are very closely tied

[7] 'Think of the architecture of our beliefs as a three-story building. At the top level are detached files… At the bottom level are perceptions and perceptual buffers. Buffers are new notions associated with the perceptions and used to temporarily store ideas we gain from the perceptions until we can identify the individual, or form a permanent detached notion for him, or forget about him. The middle level is full of informational wiring. Sockets dangle down from above, and plugs stick up from below' (Perry 2001a, pp. 120–121).

to specific ER relations since they can only host information derived on the basis of these relations. This constraint is relaxed in conceptual files. Still, conceptual files remain closely tied to ER relations in the sense that their existence (like that of proto-files) is conditional upon the existence of specific ER relations to the referent. This second constraint is relaxed in encyclopedia entries, since they do not depend upon specific ER relations for their very existence, but only on there being some ER relation or other to the referent. Still, all types of file, including encyclopedia entries, are based upon ER relations to the object they are about, and their reference depends upon the ER relations on which they are based.

27.4 Objections and Responses

27.4.1 The Circularity Objection

Two of my PhD students, Michael Murez and Gregory Bochner, have argued that the mental-file account of *de jure* coreference is circular (Murez 2009; Bochner 2010).[8] Let me start by quoting a passage from Bochner's paper:

> Many advocates of mental files, while presenting their view, themselves acknowledge—as if this were compatible with what they claim—that a mental file is created when an object is *taken to be* one by the subject.[9] (…) But to claim that a file is created for what is *(perhaps mistakenly) taken to be* an object is just to acknowledge that co-reference *de jure* rests on something akin to a prior judgement. If you already need to think of the object *in order to* determine that it is a single object deserving a single location in your syntax, then this means that you must be able think of the object *prior* to the attribution of a vehicle or mental file. And, presumably, if some identity mistake is made in this early process of syntactic assignment—if, for instance, two different vehicles are created for a unique object taken to be two distinct objects—it will be *that* early mistake that will explain cognitive significance, not the fact that there are two vehicles. (…) All of this is incompatible with the idea… that it is differences in syntax that determine differences in cognitive significance, and, instead, squarely supports the opposite view that it is differences in cognitive significance that determine differences in syntax.

[8] The circularity worry is already expressed in this passage from Lawlor's dissertation: 'In the context of building from the ground up an account of what constitutes coreferential thinking… [the theorist] owe[s] an account of what makes information belong to a single file. And [she] cannot provide this account in terms of a thinker's capacity for coreferential thinking. That would be viciously circular' (Lawlor 2001, p. 80).

[9] At this point, Bochner quotes Forbes, who writes: 'When we receive what we take to be *de re* information which we have an interest in retaining, our operating system may create a locus, or dossier, where such information is held; and any further information which we take to be about the same object can be filed along with the information about it we already possess. […] The role of a name is to identify a file for a particular object—as I shall put it, we use names to "label" dossiers. In sum, then, on coming across a new name, one which is taken to stand for some particular individual, the system creates a dossier labeled with that name and puts those classified conditions into it which are associated with the name' (Forbes 1990, p. 538; Bochner's emphasis).

As I understand it, the objection targets both the mental-file account of *de jure* coreference, and the mental-file account of identity judgements. According to the mental-file account as I have presented it, there is *de jure* coreference when two pieces of information occur in the same file and are 'presumed' to be about the same object; while identity judgements have the effect of linking two distinct files, thereby establishing *de facto* coreference between the pieces of information in the respective files (e.g. 'is bald' and 'is well read' in our earlier example). But if *de jure* coreference is a matter of belonging to a single file, and if, in turn, belonging to a single file is a matter of 'being taken to be about the same object', then (a) *de jure* coreference rests on the judgments of identity and therefore reduces to *de facto* coreference, and (b) the account of *de facto* coreference (judgments of identity) in terms of operation on files leads us into a regress, since the files themselves presuppose identity judgments: What goes into the file is whatever information we gain concerning *the same object as* information already in the file.

In response, let me start by noting that the central idea that various pieces of information cluster into a single file, when they are 'taken to concern the same object', can be understood in a way that does *not* presuppose a prior identity judgment. It may be entirely a matter of subpersonal binding of information. Thus, in the case of proto-files at least it is the cognitive system, not the subject, that takes the pieces of information to concern the same object and cluster them within a file. The subject does not judge that the pieces of information concern the same object. Identity is presupposed, it is built into the way the information is (subpersonally) packaged. So this is very different from an identity judgment, and coreference *de jure* does not reduce to coreference *de facto*.

As Murez makes clear, however, the difficulty comes from the introduction of *conceptual* files (the sort of files which are thought constituents and which are exercised when we make judgments). Something goes into a (conceptual) file if it is *judged* to concern the same individual as information in the file. The circularity objection therefore holds against what Murez calls the 'sophisticated theory' of mental files, i.e. the theory that goes beyond proto-files and makes room for conceptual files (which Murez refers to as 'relation-independent files'). Judgments of identity are accounted for in terms of a certain operation on such files (linking), yet the files in question, because they are conceptual, constitutively depend upon certain identity judgments: What goes into the file is whatever information we gain concerning (what we take to be) *the same object as* the information already in the file. But if conceptual files themselves depend upon identity judgments (in order to be fed the 'alien information' need to qualify as conceptual files), then we *cannot* analyse identity judgments in general in terms of a linking operation on files, as I have done, without launching a regress.

27.4.2 Proto-Linking

To avoid the regress, we must take advantage of the hierarchy of files and the distinction of levels it is based on. Proto-files, as Murez acknowledges, are characterized in terms of the ER relation which determines which information is allowed into

the file's content, independently of any identity judgment. So, what we must do is introduce a clear distinction between linking as it operates on *these* files (the proto-files), thereby making it possible for them to achieve status of the conceptual files, and linking as it operates on the conceptual files that result from the prior linking operation. The prior operation we can refer to as 'proto-linking' (it corresponds to what, earlier, I called 'linking-*cum*-expansion'). This distinction provides a way to avoid the regress. Identity judgments are accounted for in terms of an operation on (conceptual) files, namely linking. The very notion of a conceptual file itself presupposes some linking operation, but there is no circularity because the linking operation which the notion of a conceptual file presupposes—proto-linking—is not quite the linking operation which accounts for identity judgments. It operates on proto-files while linking operates on conceptual files; and it expands the proto-file beyond (what becomes) the nucleus while linking does not affect the structure of the conceptual files it operates on.

Of course, there is a sense in which it is the same operation of linking in both cases, namely the operation that connects two files and makes it possible for information to flow freely between them; but in one case it operates on proto-files, forcing alien information into them and expanding them accordingly, while in the other case it operates on already constituted conceptual files with a dual structure (nucleus + periphery) which it exploits but does not affect or modify. That difference between the two cases is sufficient to meet the objection. The files on which proto-linking operates are proto-files whose proper functioning does *not* rest on identity judgments; hence, no circularity is involved in analysing identity judgments as establishing a link between two conceptual files, which themselves are analysed as resulting from a (proto-)linking operation on proto-files.[10]

27.4.3 The Transitivity Objection

In a couple of recent papers (Pinillos 2009, 2011), Angel Pinillos argues that 'third object' accounts of *de jure* coreference (of which the mental-file account is an

[10] Earlier, I mentioned recognition as a rather fundamental cognitive phenomenon involving linking. Presumably, recognition exists in nonconceptual creatures, and if it does, it is an instance of proto-linking. This suggests that proto-linking by itself is not sufficient to turn proto-files into conceptual files; for if it were, it would be incoherent to assume that nonconceptual creatures can have recognitional capacities (since proto-linking would automatically endow them with concepts). There are, indeed, several reasons for decoupling proto-linking from conceptualisation. First, the limited intake of alien information that comes with proto-linking is likely to be only a first step towards satisfying the full-blooded Generality Constraint. Second, more constraints than the Generality Constraint presumably have to be satisfied for conceptual thought to emerge. Instead of assuming that proto-linking automatically converts proto-files into conceptual files, therefore, we should enrich the hierarchy of files with an extra level: Between the proto-files and the conceptual files, we should posit an intermediate category, namely the expanded proto-files which result from proto-linking. Such files remain nonconceptual yet they are hospitable to alien information (to some extent at least) and go some way towards satisfying the Generality Constraint.

instance) are bound to fail. According to such accounts, Pinillos says, two terms are coreferential *de jure* if and only if they are associated with a single entity (e.g. a single mental file) which constitutes or determines their shared cognitive significance. (That entity is the 'third object'.) Pinillos's alleged knock-down objection to such accounts is that being coreferential *de jure* is not a transitive relation. It is possible for *A* and *B,* and for *B* and *C,* to be coreferential *de jure*, even though *A* and *C* are not. But if the relation of *de jure* coreference rested on the identity of the mental files respectively associated with each of the terms, it should be transitive, since identity is a transitive relation. Pinillos takes his argument to support Fine's 'relationist' approach to *de jure* coreference (Fine 2007).

Pinillos gives examples like the following to show that *de jure* coreference is not a transitive relation:

0. We were debating whether to investigate both Hesperus_1 and Phosphorus_2; but when we got evidence of their true identity, we immediately sent probes $\text{there}_{1,2}$.
1. As a matter of fact, my neighbor John_1 is Professor Smith_2, you will get to meet (the real) $\text{John Smith}_{1,2}$ tonight.
2. Hesperus_1 is Phosphorus_2 after all, so Hesperus-slash-$\text{Phosphorus}_{1,2}$ must be a very rich planet.

He argues that in each example, there are three terms *A, B* and *C* such that *A* and *B* are coreferential *de jure*, *B* and *C* also are coreferential *de jure*, yet *A* and *C* are only coreferential *de facto*. In (0) and (2), 'Hesperus' and 'Phosphorus' are coreferential *de facto* (as they feature in the informative identity judgment 'Hesperus is Phosphorus'), yet both 'Hesperus' and 'Phosphorus' are *de jure* coreferential with 'Hesperus/Phosphorus' in (2) or with the anaphoric 'there' in (0). As Pinillos puts it,

> 'There' and 'Hesperus' are *de jure* coreferential because… one who understands the use of (0) must know that if 'there' refers at all it must corefer with 'Hesperus'. Hearers who understand the use of (0) know that those occurrences can't refer to different objects. Similarly, hearers who understand the use of (0) must also know that if 'there' refers at all, it must corefer with 'Phosphorus'. Hearers must know that they can't refer to different things. Hence, (…) 'there' is *de jure* coreferential with both 'Hesperus' and 'Phosphorus'. (Pinillos 2009)

The criterion for *de jure* coreference which Pinillos appeals to here is the knowledge criterion: When two terms are *de jure* coreferential, one cannot understand the utterance without knowing that the two terms corefer (if they refer at all).

So, example (0) shows that *de jure* coreference is not a transitive relation. Pinillos concludes that third objects account fail. According to such accounts, he says,

> Occurrences A and B in a discourse are *de jure* coreferential because they stand in a certain relation R to a single object X (e.g. a single mental file).

This cannot be right, Pinillos points out, for if it were—if *de jure* coreference was a matter of identity (identity of mental file, say)—it would be a transitive relation; but example (0) establishes that it is not.

27.4.4 Reply to Pinillos

I agree with Pinillos that example (0) shows that *de jure* coreference, as characterized through the knowledge criterion, is not a transitive relation; but I deny that this argues against the mental-file account (or third-object accounts in general). It does argue against a strong version of such accounts, namely:

> Occurrences *A* and *B* in a discourse are *de jure* coreferential (i.e. pass the knowledge test) *just in case* they stand in a certain relation *R* to a single object *X* (e.g. a single mental file).

But it does not argue against a weaker version of the theory, according to which being associated with a single mental file (or any relevant third object) is sufficient, though not necessary, for *de jure* coreference:

> Occurrences *A* and *B* in a discourse are *de jure* coreferential if they stand in a certain relation *R* to a single object *X* (e.g. a single mental file).

This weaker version leaves open the possiblity that two occurrences *A* and *B* might be *de jure* coreferential (i.e. pass the knowledge test) for some other reason than their being associated with the same mental file.

The knowledge test characterizes *de jure* coreference in terms of a priori knowledge of coreference. Now, if two terms have the same sense, understanding the terms (knowing their sense) entails knowing that they corefer; that follows from the constraint that sense determines reference. So, identity of sense entails *de jure* coreference. But why should the entailment be bidirectional? There may be other sources of a priori knowledge of coreference than sense sharing. Indeed, I will argue, that is exactly what is going on in Pinillos's examples: A priori knowledge of coreference is secured, but the source of such knowledge is not the identity of the associated mental files (or the identity of sense, more generally).

In each example, we find that two terms (say, 'Hesperus' and 'Phosphorus') are associated with distinct files—so they are only *de facto* coreferential—but there is also a third file (say the 'Venus' file or, better, the 'Hesperus/Phosphorus' file) which is created when one learns that the two terms actually corefer. That third file is what the 'merge' model posits: It says that, upon understanding and accepting an identity, one feeds all the information from the two initial files into a third file and suppresses the initial files.[11] On the weaker 'link' model, one does not (automatically) suppress the initial files, but that does not prevent one from opening a file for the unique object which is the referent of the two initial files. It is the function of slash-terms such as 'Hesperus/Phosophorus' to be associated with such files. On the 'link' model, it is possible to construe such files as including each of the initial files (now connected though linking) as subfiles. Be that as it may, given the way the inclusive file is introduced and its role, it is a priori that it corefers with each of the

[11] 'On receipt of an identity-statement invoking two…clusters, the two appropriate cards are withdrawn and a new card is prepared, bearing both the names of which one heads one of the original cards and one the other, and incorporating the sum of the information contained in the original cards; the single new card is returned to stock and the original cards are thrown away' (Strawson 1974, p. 56).

initial files. So, it is a priori that 'Hesperus' and 'Hesperus/Phosphorus' corefer (if they refer at all). Likewise, it is a priori that 'Phosphorus' and 'Hesperus/Phosphorus' corefer (if they refer at all). It follows that both 'Hesperus' and 'Phosphorus' are *de jure* coreferential with 'Hesperus/Phosphorus'. Yet these terms are *not* associated with the same file: 'Hesperus' and 'Phosphorus' are associated with what I called the 'initial files' while 'Hesperus/Phosphorus' is associated with what I called the 'inclusive file'.[12] In this case, therefore, we have an instance of *de jure* coreference that is not accounted for in terms of a shared file, but in terms of a relation other than identity between two distinct files: the relation—whatever it is exactly—that holds between the initial files and the inclusive file. Such cases can be accounted for within the mental-file account, so they do not argue against it; they only argue against an implausibly strong version resting on the (unargued) premise that only the identity of associated files can be responsible for a priori knowledge of coreference.

Acknowledgments This chapter was my Gareth Evans Memorial Lecture (Oxford, 25 January 2011); I am happy to dedicate it to Kevin for his birthday. The research leading to the lecture has received funding from ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL

References

Bach K (1986) Thought and object : *de re* representations and relations. In: Brand M, Harnish RM (eds) The representation of knowledge and belief. The University of Arizona Press, Tucson, pp 187–218

Bach K (1987) Thought and reference. Clarendon Press, Oxford

Bochner G (2010) Cognitive significance and non-descriptive senses. Ms

Campbell J (1987) Is sense transparent? *Proceedings of the Aristotelian Society* 88:273–292

Campbell J (1994) Past, space and self. The MIT Press, Cambridge

Campbell J (2002) Reference and consciousness. Oxford University Press, Oxford

Campbell J (2006) Sortals and the binding problem. In: McBride F (ed) *Identity and modality*. Clarendon Press, Oxford, pp 203–218

Castañeda HN (1999) The phenomeno-logic of the I: essays on self-consciousness. Edited by Hart J, Kapitan T. Indiana University, Bloomington

Donnellan K (1970) Proper names and identifying descriptions. *Synthese* 21:335–358

Dretske F (1988) Explaining behavior. The MIT Press, Cambridge

Evans G (1982) The varieties of reference. Edited by McDowell J. Clarendon Press, Oxford

Fine K (2007) Semantic relationism. Blackwell, Oxford

Forbes G (1990) The Indispensability of *Sinn*. *Philos Rev* 99:535–563

[12] As Fine notes, 'it is not that the merged file represents the individual as the same as the earlier files, since that would require that the earlier files represent the individual as the same. Rather, the new file, if I choose to create it, will represent the individual as being the same as the earlier files' (Fine 2007, p. 69). In this passage in which he talks about the mental-file account, Fine seems to acknowledge that 'Hesperus' and 'Hesperus/Phosphorus' are not 'strictly coreferential' (even though they pass the 'knowledge test' he himself uses—see footnote 5). Strict coreference, thus understood (i.e. narrowly), corresponds to the idea that the two terms are associated with a single file. (Fine argues against accounts of strict coreference in terms of either mental files or nondescriptive senses, but his arguments do not convince me, and, as far as I can tell, his criticism does not apply to the present account).

Geach P (1972) Logic matters. Blackwell, Oxford
Kripke S (1980) Naming and necessity. Blackwell, Oxford
Lawlor K (2001) New thoughts about old things. Garland, New York
Millikan R (1997) Images of identity. *Mind* 106:499–519
Millikan R (2000) On clear and confused ideas. Cambridge University Press, Cambridge
Murez M (2009) Mental files and coreference. In: Murez M (ed) *Self-location without mental files*. Institut Jean-Nicod, Paris, pp 47–78
Peacocke C (2012) Subjects, consciousness and perceptual content. In: Coliva A (ed) *Self and self-knowledge*. Oxford University Press, Oxford
Perry J (2000) The problem of the essential indexical and other essays, 2nd edn. CSLI Publications, Stanford
Perry J (2001a) Knowledge, possibility and consciousness. The MIT Press, Cambridge
Perry J (2001b) Reference and reflexivity. CSLI Publications, Stanford
Perry J (2002) Identity, personal identity, and the self. Hackett, Indianapolis
Pinillos A (2009) *De jure* coreference and transitivity. Ms
Pinillos A (2011) Coreference and meaning. *Philos Stud* 184: 301–324
Pylyshyn Z (2007) Things and places: how the mind connects to the world. The MIT Press/ Bradford Books, Cambridge
Recanati F (1993) Direct reference. Blackwell, Oxford
Recanati F (2010) Singular thought: in defence of acquaintance. In: Jeshion R (ed) *New essays on singular thought*. Clarendon Press, Oxford, pp 141–189
Reichenbach H (1947) Elements of symbolic logic. Macmillan, London
Sainsbury M (2002) Departing from Frege. Routledge, London
Sainsbury M (2005) Reference without referents. Clarendon Press, Oxford
Schroeter L (2008) Why be an anti-individualist? *Philos Phenomen Res* 77:105–141
Searle J (1983) Intentionality. Cambridge University Press, Cambridge
Strawson P (1974) Subject and predicate in logic and grammar. Methuen, London

Chapter 28
Did *Madagascar* Undergo a Change in Referent?

Marco Santambrogio

Abstract Kevin Mulligan has defended the view that perception is necessary for proper names to refer to spatiotemporal objects and, if the disjunctivist account of perceptual content is accepted, then the category of object-dependent singular terms must also be accepted, and proper names belong to it. I intend to take a different, more direct, route to reach the conclusion that proper names are object dependent. To ask whether, e.g., if Nixon had not existed, the name *Nixon* would have existed or whether the same name could have undergone a change in referent, is to ask whether it is a mere accident that the name *Nixon* as we have it now, names *Nixon*. Could our name *Nixon* have named, say, David Kaplan, either because it was originally given to him or because, at a certain stage, it changed its referent? This is what the issue of object dependence, as I understand it, amounts to. Here, I only address one half of the problem, namely the possibility of a change in referent. A familiar case of apparent change is that of the name *Madagascar,* first brought to our attention by Gareth Evans, and then briefly discussed by Saul Kripke in *Naming and Necessity*. Was it a real change in referent or was a new name created? I claim that the latter is the case. The chapter is mainly intended to clarify the *Madagascar* example, without drawing any general conclusion. But the issue has some obvious bearing on a number of well-known problems, including Kripke's puzzles about belief and linguistic consumerism.

Keywords Object-dependence · Change in reference · Direct reference · Madagascar case · Speaker's reference

28.1

With admirable clarity and his superb mastery of the Austro–German philosophical literature, Kevin Mulligan has presented, and developed, the view that perception and behavior are part of a language—a view that he traces back to that tradition. He has also shown how close this view comes to Kripke's account of reference (see particularly Mulligan 1997).

M. Santambrogio (✉)
Università di Parma, Parma, Italy
e-mail: marco.santambrogio@unipr.it

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_28,

The view is articulated in a number of theses. One concerns the acquisition of language. In Bühler's words, "Every speaker has gathered the meaning of all naming words from things and states of affairs pointed out directly or indirectly and then retained it in practice" (Bühler 1990).

Another bold thesis, which runs against the received dogma that only in the context of a sentence can the names function as names, says that the genesis of reference is rooted in two basic, and neglected, types of names that are used outside any sentential context: tags, which involve perception of a bearer, and signals, which involve both perception and behavior. Even in the case of names introduced by non-referential definite descriptions, a speaker could not understand them unless he possessed the ability to use tags, rooted in his perception of what they name.

I shall concentrate on the main thesis, defended by Kevin and rejected by Husserl and Frege, among others, according to which not only is perception involved in all the main types of direct singular reference but it plays a more than incidental role: it fixes reference. Reference is parasitic on perception to such an extent that names and other devices of direct reference are object dependent. I cannot here go into the details of the subtle arguments offered by Kevin.

What does this object dependence consist in? Husserl, quoted by Kevin, takes the notion as follows: "the proposition that an expression, in so far as it has meaning, relates to an object, can be interpreted in a proper or authentic sense in which it includes the object's existence. Then an expression has meaning if an object corresponding to it exists, and is meaningless if such an object does not exist" (Husserl 1970, I, § 15). It is not clear—Kevin comments—whether he intends "includes" to be used in the strong modal sense, which he rejects in any case. The strong modal sense endorsed by Kevin is (if I understand him correctly) that a directly referring singular term would not exist, or would not be the same term, if its referent were not there.[1] It is advisable, I think, to add that if the term, at its creation, referred to a given object, it cannot undergo a change in referent over time.

Kevin has offered an "unfamiliar route" for defending the view that directly referring terms are object dependent in the strong sense. It involves establishing whether or not the correct account of perceptual content is disjunctivism, i.e., the view that there is no lowest type of perceptual content such that some of its tokens are veridical contents and others non-veridical. "If object-dependent perceptual content completes the use of a demonstrative term then the latter inherits object-dependence from the [former]. If the completor is a perceptual illusion then there is no singular reference, the expression employed merely appears to be a singular term, just as the subject only appears to enjoy veridical perception. Thus we arrive at a view of referring expressions that resembles Kripke's account of rigid designators" (Mulligan 1997, p. 8).

In this chapter, I will explore a different route to reach the conclusion that proper names (not any kind of directly referring terms) are object dependent. The first step to take is to note that object dependence is not the same as rigidity. As the notion was introduced by Kripke, a designator rigidly designates a certain object if it

[1] For a systematic treatment of the notion of modal dependence, see Mulligan and Smith (1986).

designates that object wherever the objects exists, i.e., designates the same object in every possible world in which that object exists and designates nothing elsewhere. Proper names like *Nixon* Kripke maintains, are rigid designators. Of course we do not require that the objects named exist in all possible worlds. If the object designated is a necessary existent, the name is strongly rigid. Clearly, *Nixon* is not strongly rigid. There is an important, and subtle, controversy concerning the idea that a proper name would designate nothing if the bearer of the name were not to exist.[2] This, however, has little to do with the issue of object dependence, which is an ontological question about the essence of names. To ask whether, e.g., if Nixon had not existed, the name *Nixon* would not have existed or whether the same name could have undergone a change in referent is to ask whether it is a mere accident that the name *Nixon* as we have it now, names *Nixon*. Of course, *Nixon* might never have been called *Nixon* (he might never have been named at all) and someone else might have been so named. This is obvious. But could *our* name *Nixon* have named, say, David Kaplan, either because it was originally given to him or because, at a certain stage, it changed its referent? This is what the issue of object dependence, as I understand it, amounts to. The assumption that names are rigid is clearly insufficient to answer it one way or the other.

One of the leading direct reference theorists, David Kaplan, who upholds a strong view of rigidity, i.e., that names have the same designation in *all* possible worlds, takes what we have called object dependence to be an open problem. "The question, 'Is it possible that a name which in fact names a given individual, might have named a different individual?' is, for me, a substantial metaphysical question about the essence of a common currency name…. [T]here is not, I believe, an *obviously* correct answer" (Kaplan 1990, pp. 118–119). This clearly confirms, if any such confirmation were needed, that the issue of object dependence is utterly distinct from that of the rigidity of names.

The former issue is also independent from descriptivism, which is the view that senses or any kind of descriptive contents that are semantically relevant are invariably associated with names and other singular terms. As Kevin points out, Frege himself gave an elegant account of the way indexical expressions refer that, at least from the interpretation presented by Wolfgang Künne, involves a strong form of object dependence, even in the case of clearly descriptive indexical expressions such as "that table." Künne calls this type of proper names introduced by Frege, *hybrid proper names*. They contain a verbal and a non-verbal part. Even though Frege does not explicitly claim that the non-verbal parts of such names are or contain their referents, "this–says Kevin–does seem to be suggested." If a name contains its referent as its part, then it is hard to see how it could exist if the referent did not exist (and how it could stay the same if the referent changed).

In what follows, I will not take up the whole issue of object dependence but confine myself to the question of whether a name can undergo a change in referent. This addresses only one half of the problem, the other half being whether a name

[2] It is well known that some direct reference theorists, Kaplan for one, think that Kripke has misdescribed his own concept (Kaplan 1989a, p. 493).

such as *Nixon* at its creation, could have named, say, David Kaplan. Although they are related, the two questions are not to be confused.[3]

A familiar case of apparent change is that of the name *Madagascar,* first brought to our attention by Gareth Evans (1973), and then briefly discussed by Saul Kripke in *Naming and Necessity*. Was it a real change in referent or was a new name created? I claim that the latter is the case. The chapter is mainly intended to clarify the *Madagascar* example, without drawing any general conclusion. But the issue has some obvious bearing on a number of well-known problems, including Kripke's puzzles about belief. Linguistic consumerism—i.e., the idea that "[i]n our culture the role of language creators is largely reserved to parents, scientists, and headline writers for *Variety*; it is by no means the typical use of language" (Kaplan 1989b, p. 602)—is also at stake.

28.2

The problem of whether a name can undergo a change in referent was first raised by David Kaplan (as far as I know) in the last but one paragraph of "Words" (Kaplan 1990):

> can a common currency name undergo a change in referent? There is no prima facie reason against it. I re-emphasize: The identity of a common currency word lies in its continuity, both interpersonal and intrapersonal, as has been discussed. It is a matter for further analysis to say whether such an entity could change meaning (or reference). It is certainly no part of my conception that it cannot. The matter does, however, call for careful thought. One might consider two kinds of polar cases: In one case you intend to use (to repeat) a given common currency name with whatever referent it may have. ("What is Hesperus?" you ask, overhearing a conversation in which the name is used). In the antipodal case, you intend to dub a particular thing using an apt generic name. In the former case there is continuity, in the latter, creativity, a new name is created. But there are those troubling cases (first thrust upon our consciousness by Keith Donnellan, and then Gricefully reconceptualized by Saul Kripke) that seem to lie in between: the man with the Martini, the false introduction, and their ilk. (Kaplan 1990, pp. 117–118)

Before considering the "troubling cases," it is appropriate to clarify some simpler ones, in which there seems to be no denying that a new name is created. What are the prototypical cases of creation of a new name? Surely, the baptism of the Evening Star by some Babylonian is a case in point. Kaplan imagines the following story: "I imagine that at some point some Babylonian looked up in the sky one evening and said (in Babylonian) 'Oh, there's a beauty. Let us call it—' and then he introduced the name. What he did was to create a word" (Kaplan 1990, p. 100). Let us suppose that nothing was called "—" before.

Now, consider a slightly different case. Someone dubs his pet aardvark *Napoleon*. Clearly, this is *not* a case where the name of the French emperor is merely *repeated* —which is "the first kind of polar cases"— as happens when, e.g., a schoolboy

[3] The distinction is clearly made by Kripke (1980, p. 114, note 57).

hears his teacher saying that Napoleon did this and that, and then repeats "Napoleon was defeated at Waterloo." However, once it is admitted that no mere repetition is involved in the aardvark case, how do we know that, instead of Bonaparte's name undergoing a change in referent, a new name has in fact been created?

There are both differences and similarities between the two cases concerning *Napoleon*. For one thing, the aardvark's owner, unlike the student, has the *intention of not using* the name *Napoleon* with the same reference as the man from whom he heard it (whether or not he also intended so to use it when he first mastered the emperor's name).[4] For another, he clearly intends to *dub* his aardvark.

The similarity lies in the fact that the name *Napoleon* was already in use, before the dubbing occurred (unlike the *Hesperus* case, as we imagined). So, it must at least be argued that the owner is using an apt *generic* name and creating a new *common currency name,* instead of merely changing the referent of an extant common currency name. Kaplan claims that "[t]he identity of a common currency word lies in its continuity, both interpersonal and intrapersonal." But it can be supposed that there *is* continuity between the aardvark's and Bonaparte's names. Possibly, the aardvark was dubbed *in honor* of Bonaparte and it is unlikely that the latter's name was unknown to the owner.

However, it is far preferable to say that a new name is created, instead of supposing that one and the same common currency name has undergone a change in referent, the reason being that the two names coexist and can be used together, even by the aardvark's owner himself, as in, e.g., "Napoleon is not such a good strategist as Napoleon, but it occasionally catches a mouse" or, more simply, "Napoleon is not Napoleon." It would be bizarre to suppose that there is only one common currency name undergoing a change in referent in the middle of an utterance.

28.3

I now consider some cases where, without any proper ceremony of baptism taking place, a new name is created. These cases depend on the dynamics of the evolution of language and the interaction between individual speakers and the community at large.

[4] It is doubtful that every time a name is repeated, the speaker must have the intention to use it with the same reference as the man from whom he heard it. Michael Devitt (1981) pointed out that Kripke requires this intention to be contemporary with learning the name, but he does not require later uses of the name to be governed by an intention looking backward to the acquisition of the name. In any case, even if a speaker has the intention of preserving the reference of a name when learning it, this hardly prevents his subsequent using it to dub another individual. Note that, if it can be shown—as I intend to—that a name cannot change its referent, which is thus inseparable from it, the intention to use an extant name is ipso facto the intention to use it with the referent it in fact has. Whenever a change in referent occurs, the speaker is either helping herself to a different name or creating a new one. Later on, I shall point that, in some cases, it might not be entirely transparent to the speaker which name she is in fact using.

There is no denying that Kripke's notion of causal chain amounts to a *social* view of the functioning of language. Whenever a speaker uses an extant name, the reference of the name in his or her utterance is fixed by the causal chain to which it belongs. The individual speaker can rely on the community, in which the name already has its semantic value, i.e., its reference, provided that he or she intends to use the name without substantially altering it. Causal chains reach back to objects in the past but the view of language emerging from Kripke's work is still generally static, in so far as it does not consider how uses of a name within a community and by individual speakers can change over time, and how individual uses can dramatically affect conventions in the whole community. The cases I intend to consider, in which a name is created without individual speakers being aware of what is happening, are typically those where someone makes a mistake as to the correct use of a name, and then the mistake catches on and spreads to the community. Such cases, in which "what was originally a mere speaker's reference may, if it becomes habitual in a community, evolve into a semantic reference," were clearly envisaged by Kripke (Kripke 1977, p. 271). In them, a new name, a phonograph of the old one, is created. The study of such cases belongs to the "dynamic account of the evolution of language."

Strictly speaking, even clear cases of a creation of a new name by means of an explicit baptism involve some interaction between the author of the baptism and the community, even though it appears that it all depends on the intentions of the former. The interaction may be minimal and very rudimentary. For instance, when the aardvark's owner dubs her pet *Napoleon* before the new name comes into existence someone else must come to know of her dubbing. Otherwise, if the owner keeps her decision entirely to herself, no one, not even the owner herself, could possibly use the name properly in order to make herself understood. At the very least, it has to be known that the name is a name, not a meaningless sound. Unless it is circulated in the community, the name has speaker's reference, if it has any at all, not yet semantic reference.

The notions of *speaker's referent* and *semantic referent* are used here as Kripke introduced them in his 1977 article "Speaker's Reference and Semantic Reference." I quote: "The speaker's referent is the thing the speaker referred to by the designator, though it may not be the referent of the designator, in his idiolect" (Kripke 1977, p. 264). As to semantic reference: "If a speaker has a designator in his idiolect, certain conventions of his idiolect, (given various facts about the world) determine the reference in the idiolect: that I call the semantic referent of the designator" (Kripke 1977, p. 262).

Let us consider some cases in which a speaker makes a particular kind of mistake in using a name. The first case is almost the same as one envisaged by Kripke. Two people, *A* and *B,* see Smith in the distance. Mistaking him for Jones, *A* says "Jones is raking the leaves."[5] As Kripke has gricefully reconceptualized the case,

[5] The original example is as follows: "Two people see Smith in the distance and mistake him for Jones. They have a brief colloquy: 'What is Jones doing?' 'Raking the leaves'" (Kripke 1977, p. 263). In my version, it is, crucially, only *A* who mistakes Smith for Jones.

A has a *specific* intention to refer to the man in the distance, who is in fact Smith, a false belief that the person fulfills the conditions for being the semantic referent of the name she uses, and also a *general* intention to refer to the semantic referent of the name. Due to the fact that *A* has more than one intention concerning the use of the name *Jones,* the case is *complex* and *A's* use is *referential*. Even though *A* has said something *of* Smith, the name *Jones* still has Jones as semantic referent. Smith is but the speaker's referent. This does not affect the language community and, in particular, the name's semantic referent.

Now, suppose that *A's* utterance is part of a conversation with *B*. Under certain conditions, now to be investigated, it is possible that *A* may affect the community's usage, in so far as the speaker's referent of the name *Jones* evolves into its semantic referent within the restricted community formed by *A* and *B* alone. It can then be claimed that a new name is created.

How can *B* react to *A's* utterance? Various possibilities come to mind. First, *B* might realize that *A* has made a mistake and say, e.g., "Look, the man raking the leaves isn't Jones. Jones is at home watching TV." Here, *B* is using the name *Jones* with the semantic referent it has within the community to which both he and *A* belong. What he intends to point out is that the speaker's referent and the semantic referent are distinct.

Second, *even though he realizes A's mistake, B* decides to follow suit and says, e.g., "Jones looks tired. He has been working hard." Here, *B* uses the name just as *A* does, even though he himself makes no error of misidentification.

Third, *B* mistakes Smith for Jones, just as *A* is doing, and the conversation continues with both *A* and *B* using *Jones* when they want to say something about Smith. (This case is similar to the original one considered by Kripke).

In all cases, *B's* reactions are perfectly natural and appropriate to the circumstances. In the first and the third cases, nothing happens concerning the reference (either the speaker's or semantic) of the name *Jones*. Let us consider more closely the second case. *B* realizes *A's* mistake but for some reason (he might be too lazy to correct *A* or just not interested, or he might have other motivations[6]) he follows suit. Suppose that no one, besides *A* and *B,* is present. Then, *B* must intend to make himself understood by *A,* since there would be no point in fooling *A* by saying what she cannot understand. Clearly, *B's* intention, in using the name *Jones,* is to refer to the same person as *A*. But *A* had two distinct intentions: Her speaker's referent was Smith, whereas the semantic referent was Jones. Which one does *B* intend to refer to? For sure, *B* intends to refer to Smith since, otherwise, he would not have uttered what he did. The assumption that he intended to make himself understood by *A* is crucial here.

Now, what are, respectively, the speaker's referent and the semantic referent of the name *Jones,* as used by *B*? *B* is in a very different position to *A*. Like *A,* he has a *specific* intention to refer to Smith, but he knows that Smith does *not* fulfill the conditions for being the semantic referent of *Jones,* and he does not have the

[6] As in the example, considered by Donnellan (1966, p. 290), of the usurper whom everyone call "the king."

general intention to refer to the semantic referent of the name as it is used in the community at large. Still, it is not yet clear that *Jones,* as used by *B,* does not have the same speaker's referent and the same semantic referent as it has when used by *A*.

I now give two arguments to show that this is not so. To repeat, this is what semantic reference amounts to:

> If a speaker has a designator in his idiolect, certain conventions in his idiolect (given various facts about the world) determine the referent in the idiolect: that I call the *semantic referent* of the designator. (Kripke 1977, p. 263)

The term *idiolect* is used here somewhat loosely, if only because it is unclear what a convention might amount to in an idiolect proper. In a footnote, Kripke adds: "… the conventions regarding names in an idiolect usually involve the fact that the idiolect is no mere idiolect, but part of a common language, in which reference may be passed from link to link" (Kripke 1977, p. 273, note 20).

The term *convention* is also in need of some clarification. For a convention to exist, it seems to be required that all the competent speakers, or at least most of them, know something of each other and at least some reciprocity is in place. Now, the image drawn by Kripke in *Naming and Necessity* does not involve any reciprocity at all. Someone, let us say, a baby, is born; his parents call him by a certain name, and the name is spread from link to link as if by a chain. But it is quite possible that each user acquiring the name is only aware of the speaker from whom he got it and is completely in the dark concerning all other previous, and subsequent, speakers. This seems to be clearly insufficient for a convention to be established. Since Kripke is explicit that everything in his article is meant to be compatible with *any* view of proper names, in particular his own, and the image drawn in *Naming and Necessity* does not support the idea of a convention existing among the users of a name, I conclude that the term *convention* is used somewhat loosely here.

For a name to semantically refer, it would certainly be too much to require, as Evans did, that:

> 'NN' is a name of *x* if there is a community *C* in which it is common knowledge that members of *C* have in their repertoire the procedure of using 'NN' to refer to *x* (with the intention of referring to *x*). (Evans 1973, p. 18)

It is entirely possible that very few members of the intended community know anything about the reference of a name for sure. Reference is compatible with all sorts of errors being common in the community. Common knowledge of the sort envisaged by Evans can hardly obtain. Still, someone at least must be in the know. Would it be possible for *everyone* to have mistaken beliefs as to the referent of a name? Suppose a pair of identical twins exists, Peter and Paul, who were so baptized at birth and then exchanged in their cribs, without anyone being aware of it, all memory having been wiped out of the order in which the two baptisms took place, so that everyone believes that Peter is the one who was, in fact baptized, *Paul* and vice versa, and calls them accordingly. Would we say that the semantic referent of either name, unbeknownst to the whole community, is the "wrong" twin? It seems to me that we would not. Even though it is possible for a large majority of the speakers in a community, or even nearly all of them, to be mistaken as to the semantic referent

of a name, someone must know. Should no one be left who has the appropriate sort of knowledge, for any name in use it must still be possible, at least in principle, to discover what its referent is.

Now, what do *A* and *B,* in our example, believe as to the referent of *Jones,* as they have used it? *A* mistakenly believes that *Jones'* semantic referent is Smith and, from *B's* utterance, infers that *B* believes the same and also believes that he, *A,* believes it. *B,* on the other hand, knows that, not only does *A* use *Jones* to refer to Smith, but *A* also believes that he, *B,* does the same and believes that *A* so believes, etc. Even though it is not common *knowledge* that *Jones* refers to Smith, since *A's* beliefs fall far short of amounting to knowledge, there is at least a common belief that it is so. Moreover, *B knows* how things really are. This seems to be sufficient to conclude that Smith is the semantic referent of *Jones,* as used within the restricted community formed by *A* and *B* alone.

Another reason to claim that, for the community formed by *A* and *B* alone, *Jones* semantically refers to Smith is as follows. Suppose that *A* and *B* go on at great length exchanging their views about Smith raking the leaves, always using the name *Jones* for him, even after *A* realizes his initial mistake. Clearly, after a while, *Jones* has become a name of Smith or at least a nickname of him. Nicknames are supposed not to require explicit dubbing and to stick to their referent by a more gradual process. But, has anything relevant happened after the first two utterances by *A* and *B*?

Michael Devitt (1981) has claimed that *multiple grounding* is required for a name to refer semantically to its referent. Clearly, if *A* and *B* go on at great length talking about Smith and calling him *Jones,* the name *is* in fact multiply grounded. But what is the use of multiple grounding? On the face of it, it is only necessary in order to spread a name from the restricted community in which a name has originated to a wider one, since the number of those in the know as to the referent of a name must be proportional to the size of the community. It does nothing to secure the referent to the name. In the community formed by *A* and *B* only, the grounding seems to be completed after the first two utterances.

It might be supposed that it is crucial for *A* to come to realize his initial mistake. But this goes against the fact, noted above, that names are secured to their semantic referents, even if all sort of mistaken beliefs are widespread within the relevant community. I conclude that Smith, who was the mere speaker's referent of *Jones,* as used by *A,* has evolved into its semantic referent, after *B's* utterance, within the community formed by *A* and *B* alone.

It is intuitively obvious that the notion of semantic reference is relative to some linguistic community or other. This has no tendency to show, however, that any kind of relativism is in the offing, in the sense that the same utterance might be true when it is taken by the lights of one linguistic community and, at the same time, false by the lights of another. The circumstances relative to which an utterance is to be evaluated are always those of the context where it occurs—the linguistic community involved being one of the context's components. In particular, *A's* utterance above is false in its own context and it does not become true simply because the name *Jones* occurring in it acquires a new referent relative to the (*A, B*) community, which comes into existence, so to speak, only after *B's* utterance. A rather intuitive

way of characterizing what has happened is as follows. Think of the name *Jones* in *A's* utterance as an arrow. The arrow misses its target and, as a consequence, the whole utterance does as well and is therefore false. Then, *B* draws a new target around the spot where the arrow has fallen, so that it appears that the arrow has hit the target right in the center. But, of course, it matters a great deal whether the target was set in place before or after the shot.

Has the name *Jones* acquired a new referent within the (*A, B*) community? Or has a new name—a phonograph, in Kaplan's terminology—been created? Even though Smith is now named, or nicknamed, *Jones,* Jones himself has not been forgotten by *A* and *B* and, of course, has kept his name. There is nothing unusual in the fact that more than one person is called *Jones*. Their common currency names are available to be simultaneously used in the same utterance as, e.g., in "Of course Jones, over there, is not Jones, who is comfortably watching TV at home." Since it would be preposterous to say that a name changes its referent in the middle of an utterance, as we saw above, we must conclude that in the utterance two distinct names occur, one of which must therefore have been created.

As a matter of fact, this conclusion should not be surprising in the least, as it could have been reached immediately by pointing out that *B,* when he realized that *A* was making a mistake and decided to go on calling Smith *Jones* nonetheless, was in the same position as the owner of the pet aardvark, who gave it the name *Napoleon,* even though he knew that the name was already taken by the emperor. It might be objected that even for the baptism of an aardvark some measure of solemnity is appropriate, which is entirely missing in *B's* sudden compliance with *A's* example and calling Smith *Jones*. But the answer simply is that no formal public ceremony is needed. As Kaplan pointed out in "Demonstratives," "a fleeting 'Hi-ya Beautiful' incorporates all the intentional elements required for [him] to say that a dubbing has taken place" (Kaplan 1989a, p. 560).[7] In any case, the more roundabout path we followed to reach the conclusion that a new name was created, was not entirely useless, as we shall now see.

28.4

The first example that comes to mind of a name that *would* have changed reference *if* it had stayed the same name, is that of *Madagascar*. According to Evans, the name was originally used in West Africa to refer to a region on the continent:

> In the case of 'Madagascar' a hearsay report of Malay or Arab sailors misunderstood by Marco Polo…has had the effect of transferring a corrupt form of the name of a portion of the African mainland to the great African Island. (Isaac Taylor, *Names and their History,* 1898). (Evans 1973, p. 11)

[7] I find it hard to reconcile this remark of Kaplan's with his characterization of consumerism quoted above, from "Afterthoughts."

We can assume that for a long time after Marco Polo no other European visited either the great African island or the region on the mainland. No one in Europe had any further contact with either the African natives or the Malay and Arab sailors. Everyone in Europe entirely relied on Polo and used the name *deferentially* with respect to him—i.e., with the overriding intention to conform to his use of it (Evans 1973, p. 21). The island, however, appeared on the maps more or less in its proper location, due to Polo's accurate descriptions in his book *The Million*. A few centuries later, Europeans began to have a number of contacts, both direct and indirect, with the island. No doubt, the name *Madagascar,* as we now use it, semantically refers to it.

Thus, *a* name *Madagascar* was used by the natives on the continent with one semantic referent and *a* name *Madagascar,* somehow causally connected with the former, is used by us, with a quite different semantic referent. Are the two names one and the same? Let it be noted that very few Europeans have realized, as Isaac Taylor did, that a change in referent (and possibly also in the name) occurred. The thing could have escaped notice altogether. For sure, Marco Polo did not realize that he had misunderstood his informers.

At least four different categories of users of *Madagascar* exist: the natives on the continent, Marco Polo, the Europeans who read *The Million* shortly after it was published, and us, a few centuries later. The first thing to establish is, when exactly did the change in reference occur? We can be assured that Marco Polo, in using the name *Madagascar* in *The Million,* had the general intention to refer to the semantic referent of it, as it was used by his informants, since the book aims at giving an entirely faithful report of the countries he had visited. There is no fiction in the book, nor are there inaccuracies that could be avoided. Of course, all the names used in it are likely to be twisted, as they were translated from their original languages into thirteenth-century French, but that is all. Polo also harbored a false belief that the great African island fulfilled the conditions for being the semantic referent of the name, and a specific intention to refer to the island. Like speaker A, in the Smith–Jones example, he thus had two distinct referential intentions.

As a rule, when one picks up a name from someone else, one intends to make use of it with its semantic, not with its speaker's, referent, if only because he may be addressing other speakers in the community and it would be unwise to let the informant's possible misunderstanding affect one's usage. However, there are exceptions. The first readers of *The Million* are likely to be exceptions to that rule. They were using the name deferentially with respect to Marco Polo and had no chance of finding themselves addressing a member of the community where the name had originated. They had no reason to think that Marco Polo was mistaken concerning the name and, even if he was, there was no point for them in trying to straighten things out and no chance of succeeding. They were entirely dependent on Marco Polo. Thus, somehow, they were in the same position as speaker *B* in the previous example, in that they did not care about any possible divergence between the semantic and the speaker's referent (which they did not, unlike *B,* know anyway). Perhaps, speaker *B* is an exception to the rule above, too, since he was using *Jones* to refer to Smith and there is a sense in which he had picked up the name from *A*.

Since, in the restricted community formed by Marco Polo and the first readers of his book—the African natives being conspicuously absent from it—Marco Polo was the only authority entitled to give a final answer to the question "Where is Madagascar?", it can be concluded that the semantic referent of the name was what Polo took it to be. There is a close similarity between that community and the one formed by speakers *A* and *B* alone, in the previous example. The main difference between the two is that no one belonging to the former community was aware that *Madagascar* had undergone a change in referent or had been substituted with a newly created name.

If the multiple grounding does nothing to secure a referent to a name and is only needed to spread the name, with its referent, to a wider community, then we might have another argument to the same conclusion. There is no denying that we now use *Madagascar* to semantically refer to the island. We too are depending on *The Million* concerning the referent of the name. But nothing relevant has happened concerning the name, except multiple grounding, since the book was first published and read. Thus, already in the restricted community formed by Marco Polo and his first readers, the island was the semantic referent of it.

It is not unlikely that also the case imagined by Evans (and discussed by Devitt 1981) of the name *Ibn Kahn* is to be understood along the same lines. Evans imagines that a mathematical manuscript was discovered in the vicinity of the Dead Sea, in which the name *Ibn Kahn* occurs, which is generally taken to name the author of it. Mathematicians often refer to some theorem of Ibn Kahn's. Suppose that, as a matter of fact, Ibn Kahn was the scribe (or an impostor). If, in discussing the import of that theorem, a mathematician states "Here Ibn Kahn uses *reductio ad absurdum,*" is he saying something true of the author of the manuscript or something false of the scribe? The former seems to be the correct thing to say. Within the mathematical community the name *Ibn Kahn* appears to have the author as its semantic referent.[8]

Let us return to *Madagascar*. Has the name changed its referent or has a new name been created, after Marco Polo? The old name is still available, with its referent, as shown by Isaac Taylor's book. An utterance such as "Madagascar is not the island of Madagascar" is perfectly intelligible and also true. As we saw above, this

[8] The case bears some similarity to one all too briefly discussed by Kripke in *Naming and Necessity*: "More exact conditions [for reference to take place] are very complicated to give. They seem in a way somehow different in the case of a famous man and one who isn't so famous…. If … the teacher uses the name 'George Smith'—a man by that name is actually his next door neighbor—and says that George Smith first squared the circle, does it follow from this that the students have a false belief about the teacher's neighbor? The teacher doesn't tell them that Smith is his neighbor, nor does he believe Smith first squared the circle. He isn't particularly trying to get any belief about the neighbor into the students' heads. He tries to inculcate the belief that there was a man who squared the circle, but not a belief about any particular man—he just pulls out the first name that occurs to him—as it happens, he uses his neighbor's name. It doesn't seem clear in that case that the students should have a false belief about the neighbor, even though there is a causal chain going back to the neighbor (Kripke 1980, pp. 95–96)." The teacher seems to have created a new name, if only because he intended to discontinue using *George Smith* with its semantic referent, and the students defer to him.

is a powerful argument to maintain that a new name has appeared, even though no one ever realized it until Taylor discovered what had happened.

28.5

None of the considerations above rested on the assumption that Kaplan's theory of common currency names holds true. The thesis that proper names do not change their referent is acceptable, it seems to me, no matter what theory of names one endorses. Kaplan, however, has additional reasons in favor of it. He distinguishes *common currency* from *generic* names. The former refer—each one to its own unique bearer. Generic names either have multiple referents or do not name anyone: "…the generic name doesn't name anyone (doesn't name any*one*, perhaps it names or is an unnatural kind). Furthermore, it doesn't pretend to name anyone (as certain common currency names do)" (Kaplan 1990, p. 111). In any case, they are quite unlike proper names. The distinction seems to be reasonably clear. However, if a common currency name were allowed to undergo a change in referent, while still being available to be used with its old referent—as, e.g., in "Madagascar is not Madagascar" and "Jones is not Jones"—it would have multiple referents, or something of the sort, and the distinction would then be blurred.

From the thesis that names can be created unwittingly, within a restricted language community, some consequences can be drawn concerning Kaplan's theory. Some commentators (e.g., Cappelen 1999; Hawthorne and Lepore 2011) maintain that intentions have a constitutive role to play concerning common currency names: "If someone intends to produce the same word *w* as that used in a particular performance, then whatever comes out of his mouth (or from his pen) is a performance of *w*" (Hawthorne and Lepore 2011, p. 461). *Intending it to be so makes it so*. They claim that the thesis—which they do not endorse anyway—is implicit in the Kaplanian notion of repetition. Let us review what we have found so far that is relevant to the issue.

We have seen that speaker *B,* in using *Jones* to refer to Smith while realizing *A's* mistake, was in a position similar to that of the man who bestows the name *Napoleon* on his pet aardvark. Even though *B* may not be fully aware of it, he has created a new name, relative to the very restricted (*A, B*) community. Since the name is new, *B* has *not* repeated the name he received from *A*. Note, however, that *B's* and *A's* names sound the same, and refer to the same referent. Also, *B* intends to use *Jones* with the same reference as *A*—that is, with the speaker's reference it had with *A*.

As to *Madagascar,* we have seen that the early readers of *The Million* were in the same position as *B* above, except that they could not be aware that a change in semantic referent had occurred with respect to a community from which the natives were absent. As before, their referent was the same as Polo's speaker's referent. However, when using *Madagascar,* they *intended to repeat* the name as used by Polo. Intending to repeat a name is no guarantee of success. Repeating a name is thus no straightforward matter. The identity of common currency names remains hard to determine.

In any case, the thesis above, that one cannot fail to produce the same name if one intends to do so, is false, as can be established by an independent argument. Suppose a speaker has heard several people refer to some Jones. As a matter of fact, some of those people were talking about one Jones, whereas others were talking about another. Our speaker misunderstood what they were saying and thought that only one Jones was mentioned. As there is no reason to suppose that one can only repeat names one has heard only once, he forms the intention to repeat "the" name. Can he succeed? Kaplan himself is adamant that, in cases such as this, "nothing whatsoever is being said. Is it [a given black box] transmitting the first word? Is it transmitting the second word? I think there is just no answer to that question" (Kaplan 1990, p. 109). In any case, the speaker fails to produce *the* name he intends to, if only because there is no such a name. (I am assuming here the thesis above only applies to common currency names. It would not be so difficult to repeat a generic name, but it would also be quite irrelevant "for serious semantics").

Be that as it may, the notion of repetition is far from being clear. Kaplan proposes a thought experiment: "Consider this thought experiment: I say the name of an individual, possibly a name known to the person to whom I am speaking. The subject has to wait for a count of five, and then repeat the name. I say a name, then the subject says the name…. So, if I say 'Rudolf,' the person says 'Rudolf'; 'Alonzo' –'Alonzo'; 'Bertrand'– 'Bertrand,' and so on" (Kaplan 1990, p. 102). Now, suppose the subject hears from Kaplan the common currency name that is Kaplan's own, *David,* but thinks that he heard David Lewis's name. When he repeats *David,* has he succeeded in producing the same name he intended to? It is not obvious that a clear answer can be given. This throws some doubt on the notion of repetition itself.

Acknowledgments I have benefited greatly from discussions with Andrea Bianchi, who has also read and commented parts of this chapter. I would like to thank Giulia Felappi for her helpful comments on an earlier version of it.

References

Bühler K (1990) Theory of language. Benjamins, Amsterdam
Cappelen H (1999) Intentions in words. NOUS 33(1):92–102
Devitt M (1981) Designation. Columbia University Press, New York
Donnellan K (1966) Reference and definite descriptions. Philos Rev 75(3):281–304
Evans G (1973) The causal theory of names. Proceedings of The Aristotelian Society: Supplementary volume 47:187–208
Hawthorne J, Lepore E (2011) On words. The Journal of Philosophy CVIII, 9:447–85
Husserl E (1970) Logical investigations. Kegan Paul, Routledge
Kaplan D (1989a) Demonstratives. In: Almog J, Perry J, Wettstein H (eds) Themes from Kaplan. Oxford University Press, Oxford
Kaplan D (1989b) Afterthoughts. In: Almog J, Perry J, Wettstein H (eds) Themes from Kaplan. Oxford University Press, Oxford

Kaplan D (1990) Words. Proceedings of The Aristotelian Society: Supplementary volume 64:93–119
Kripke S (1977) Speaker's reference and semantic reference. Midwest Stud Philos II:255–276
Kripke S (1980) Naming and necessity. Basil Blackwell, Oxford
Mulligan K (1997) How perception fixes reference. In: Burri A (ed) Language and thought. Walter de Gruyter, Berlin, pp 122–138
Mulligan K, Smith B (1986) A relational theory of the act. In: Bonomi A, Woodruff Smith D (eds) Topoi 5, Current Issues in Phenomenology. Swiss Philosophical Preprint Series, Swiss Portal for Philosophy, pp 115–130

Chapter 29
Live Metaphors

Anne Reboul

Abstract In this chapter, I outline two successive versions of the Relevance-Theoretic account of metaphors, the one initially proposed in Sperber and Wilson (Relevance: communication and cognition, 1995) and the new one recently proposed by Carston (Thoughts and utterances: the pragmatics of explicit communication, 2002) and apparently adopted by Relevance Theory. The first one claimed that metaphors have propositional effects (implicatures) while the second claims that metaphors have an explicature recovered through the construction of an ad hoc concept. Both accounts are continuous accounts (i.e., they do not posit any specific interpretation process for metaphors) and both ignore the nonpropositional (sensory) effects of metaphors. But, while the first does succeed in accounting for the propositional effects of metaphors and for the impossibility of paraphrasing live metaphors without loss, this is not the case of the second, which fails on both counts.

Keywords Live metaphor · Ad hoc concept · Implicature · Explicature · Sensory effect

29.1 Introduction

Accounts of metaphors can be distinguished on the basis of whether or not they propose that metaphors are interpreted through a specific interpretation process, different from the one used in nonmetaphorical utterances. Accounts that defend such a specific interpretation process are *discontinuity* (or *discontinuous*) accounts while accounts that reject the notion of an interpretation process specific to metaphors are *continuity* (or *continuous*) accounts.

In the present chapter, I shall mainly be interested in a specific continuity account, the account originally given of metaphor in Relevance Theory (Sperber and Wilson 1995), as well as its recent modifications as proposed by Carston (2002).

I shall begin by introducing classical accounts of metaphors in terms of *figurative meaning,* as well as discontinuist accounts building on them (e.g., Searle 1985).

A. Reboul (✉)
Laboratoire sur le Langage, le Cerveau et la Cognition (L2C2 CNRS UMR5304),
Institut des Sciences Cognitives Marc Jeannerod, 67 Bd Pinel, 69675 Bron cedex, France
e-mail: reboul@isc.cnrs.fr

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_29,

I shall show that the first such accounts meet with major difficulties, due to the fact that figurative meanings are dangerously near to paraphrases (indeed, it is difficult to see what else they could be) and that it is a well-known fact about metaphors, at least live ones, that they cannot be paraphrased without loss. One objection to the second kind of accounts is that they propose to recover figurative meanings through interpretative processes triggered by the necessary falsity of metaphors. However, not all metaphors are false and, what is more, for those that are false, negating them does not usually make them nonmetaphorical. Then, I shall outline the two successive accounts given of metaphors in Relevance Theory and show why the second account is less successful than the first while the first has nothing to say about the nonpropositional effects of metaphors (about the nonpropositional effects of metaphor, see, e.g., Davidson 1978; Guttenplan 2005).

29.2 Discontinuist Accounts

Traditional accounts of metaphor have claimed that metaphors have two meanings:

- A *literal meaning* (which is generally false)
- A *figurative meaning* (which can be true or false, but which is generally true)

Let us begin with an example1:

1. The sleep of reason begets monsters.
2. Suspending the activity of reason produces massive irrationality.

(1) is the metaphor, and if interpreted literally, it is false, probably necessarily so: reason is not the kind of entity that can sleep, neither can it beget (in the biological sense) anything. By contrast, (2), the so-called "figurative" meaning of (1), seems true (if not trivially true). What is more, some fairly recent accounts (e.g., Searle 1985) have defended the idea that it is the falsity (or conceptual incoherence) of metaphors that triggers the interpretation process through which figurative meaning is retrieved.

This approach, however, meets with two important difficulties, both to do with the notion that falsity is the central characteristic of metaphor. The first one is that all metaphors are not false, as shown by example (3)[2]:

3. No man is an island.

Another, and potentially more devastating objection, is that if metaphors *had* to be false, given the semantics of negation, (false) negated metaphors should stop being metaphors. As shown by examples (4) and (5), this is not the case:

[1] This is the English translation of the sentence Goya inscribed on the frontispiece of his *Capricios*.

[2] This sentence, from Donne's XVIIth meditation, has been widely used to prove exactly the point I am making, i.e., that all metaphors are not false (thus, this is neither a new example nor a new argument).

4. John is a bulldozer (he does not care for other people's feelings).
5. John is not a bulldozer (he is a sweet and sensitive man, considerate of other people's feelings).

The fact that metaphors can be true makes the classical position with its hypothesis of double meanings, as well as any hypothesis positing a specific interpretation process triggered by falsity, rather fragile to say the least.

There is an additional problem for such discontinuist accounts of metaphors, which is that they cannot account for a major feature of metaphor, i.e., the fact that a metaphor cannot be paraphrased without a loss. This can be seen with example (1) above, which, arguably, is a live and creative metaphor, whose paraphrase in (2) seems to lose all creativity and liveliness. In addition, the very notion of a figurative meaning seems understandable only as something that can be linguistically formulated, which strongly suggests that it is a (if not *the*) paraphrase of the metaphor. So, neither of those two discontinuist accounts can explain the impossibility of paraphrasing a metaphor, which is hardly surprising as they probably would have either to deny that metaphors cannot be paraphrased or that figurative meaning is a paraphrase of the metaphor. I shall return below to the question of paraphrase and its link with metaphor (see Sect 29.6).

29.3 Metaphors in Relevance Theory: The Original Account

In the present chapter, I shall mainly be concerned with the account of metaphor given in Relevance Theory (Sperber and Wilson 1995). This, in contrast with the analyses discussed above, is what has come to be known as a *continuity* hypothesis, the hypothesis that metaphors are interpreted just like any other utterance, or, in other words, that they do not call for a specific interpretation process. This, of course, has to depend on a rather specific view of figurative utterances, and in Relevance Theory, metaphors are treated as a species of vague communication. The idea is that all utterances are *interpretations*[3] of a thought of the speaker, but they can be more or less literal, depending on their similarity with that thought. The notion of *similarity* is, of course, a vexed one (Goodman 1970) and it was given the following meaning in Relevance Theory:

> Similarity between two propositional representations is defined as depending on the number of implications the two representations would share when interpreted relative to the same context[4].

[3] In a specific, technical, sense of the term "interpretation," which has to do with the relation between two representations, one of which is the representation (the "interpretation" in Sperber and Wilson's sense) of the other.

[4] The notion of context in Relevance Theory is not, as it is in some semantic theories, the immediate environment in which the communication is taking place. In Relevance Theory, the context is a set of mental representations in propositional forms taken from three sources: the perception of

There are then three possibilities:

- When the utterance is a *literal* interpretation of the thought, the set of implicatures of the utterance and the set of implicatures of the thought (interpreted relative to the same context) are identical.
- When the utterance is a *less than literal* interpretation of the thought, the set of implicatures of the utterance and the set of implicatures of the thought (idem) have a nonvoid intersection (or, in other words, an intersection different from the null set).
- When the utterance is not an interpretation of the thought (and there is *no resemblance* between the utterance and the thought), there is no intersection between the two sets of implicatures of the utterance and of the thought (idem; in other words, the intersection is the null set).

This means that all utterances can be localized along a continuum that goes from complete literality (a limit point) to the case in which the utterance is not an interpretation of the thought anymore (the other limit point): Most utterances (including metaphors, and more generally tropes) will come in between as less than literal interpretations of the speaker's thought.

This view of communication and utterances claims that implicatures are central to any account of metaphor, just as they are central to any account of linguistic communication. Indeed, a specificity of metaphors relative to vague utterances, according to Relevance Theory, is that they communicate weakly a potentially unlimited array of implicatures while vague utterances strongly communicate one or two implicatures. The basic idea is that a proposition is strongly implicated by an utterance if it has to be recovered for the utterance to have a relevant interpretation; it is weakly communicated if its recovery can participate in the construction of a relevant interpretation, though the utterance suggests a range of other implicatures, any of which would do as well. Thus, weakly communicating a wide array of implicatures leaves the hearer a greater liberty of interpretation than communication generally does.

In other words, metaphors communicate a wide array of implicatures none of which has to be specifically recovered for the utterance to meet the expectations of relevance while vague utterances strongly communicate an implicature that has to be specifically recovered for the utterance to meet the expectations of relevance.

Let us have a look at some examples:

6. It is 5:30.
7. Caroline is a princess.

Let us suppose that when (6) is produced, it is in fact 5:28. Most of the implicatures of *It is 5:30* are the same as those of *It is 5:28*, for instance that one should hurry up (or, alternatively, that there is no need to hurry, one is on time). (7), if said of a commoner, is presumably a metaphor. One interpretation would be that Caroline

---

the environment in which the communication is taking place; the interpretation of previous utterances in the ongoing conversation; knowledge of the world as represented in long-term memory.

is a spoilt girl, unlikely to dirty her hands to help anyone. Other interpretations are available, nevertheless, for instance that she is well dressed or elegant, etc.

A second specificity of metaphors relative to vague utterances is the reason the speaker choose to use them:

- Regarding vague utterances, the reason is one of economy: The vague utterance is sufficiently similar to the thought and is less costly to process.
- Regarding metaphors, the reason is that there was no other way of expressing the speaker's thought, because it was too complex to be expressed literally.

But the main conclusion, as said above, is that metaphors do not need a specific interpretation process: They are interpreted, just as all utterances are, through the derivation of implicatures.

So, in sum, the original account of metaphors in Relevance Theory is that metaphors weakly communicate a wide array of implicatures, which is why they cannot be paraphrased and why there cannot be a figurative meaning for a metaphor (none of its implicatures is strongly communicated enough to count as a figurative meaning). There is no literal way of expressing the speaker's thought.

29.4 Metaphors in Relevance Theory: The New Account

This was the account Relevance Theory proposed of metaphors until Carston's (2002) book. She proposed a highly different account, still claimed to be squarely in the theoretical framework of Relevance Theory, but different from the original account in some major ways. On the face of it, Carston's account[5] does keep quite a few things from the vintage account: Her's still is a continuity account (it still claims that metaphors are interpreted just as are other utterances); again, it does not explicitly reject the idea that metaphors cannot be paraphrased, nor the hypothesis that there was no other way of expressing the speaker's thought, or the idea that metaphors weakly communicate a wide array of implicatures.

There is, however, a major departure between the vintage and the current accounts: While the vintage account insisted that metaphors were interpreted just as utterances in general are because they give rise to *implicatures* (and are thus in need of pragmatic interpretative processes), the current account insists that metaphors are interpreted just as utterances in general are because they give rise to *explicatures* (and are thus again in need of pragmatic interpretative processes). Explicatures are taken to be a part of what is explicitly communicated, developments of the *logical* or *propositional form*[6] of an utterance. The pragmatic processes implicated in these

[5] I shall call the account Carston proposes the "current Relevance Theoretic account" (or the "current account"), by contrast with the original account that I shall call the "vintage Relevance Theoretic account" (or the "vintage account"), in what follows.

[6] In Relevance Theory, there is a distinction between the logical form of an utterance, which is the result of the linguistic (phonology, syntax, semantics) interpretation process, and the propositional

developments are *broadening* (widening a concept extension) and *strengthening* (reducing a concept extension).

The current view is that what happens when such pragmatic processes are operating in the interpretation of an utterance is that a concept in the propositional form (the concept that is the object of broadening or strengthening) is replaced by a so-called ad hoc concept. The notion of an ad hoc concept was introduced by Barsalou (1983), but I shall limit myself here to the account of the notion given by Carston (2002). According to that account, ad hoc concepts are constructed on the fly, through the processes of broadening and strengthening, which are themselves triggered and guided by the context and expectations of relevance. Thus, the current Relevance-Theoretic account of metaphors comes to this: *Metaphors are interpreted through the inclusion in their propositional forms of one* (*or several*) *ad hoc concepts replacing some* (*standard*) *concepts in the logical or propositional form.* This is also true of vague utterances, which explains why Carston's account is still a continuity account.

Let's begin by a vague utterance:

8. This steak is raw.
9. THIS STEAK IS RAW* [7] (RAW* = UNDERCOOKED).

The concept raw* is obtained by broadening, i.e., in this specific instance, by widening the extension of the concept raw to include undercooked dishes.

Let's now go back to our example of a metaphor:

10. Caroline is a princess.
11. CAROLINE IS A PRINCESS** (PRINCESS** = A WOMAN NOT FROM A ROYAL FAMILY, SPOILT AND SELFISH, UNLIKELY TO HELP).

The explicature in (11) is obtained by a double process of first broadening the concept PRINCESS to include commoners (i.e., the resulting ad hoc concept PRINCESS* includes all women, regardless of their family origins), and then strengthening the concept PRINCESS* by including in its definition that it denotes spoilt and selfish women, unlikely to help (here, the extension is reduced by taking a feature widely believed to be true of princesses and making it a definition of the ad hoc concept PRINCESS* yielding the final ad hoc concept PRINCESS**).

In the vintage Relevance-Theoretic account, metaphors were supposed to have two distinguishing features:

---

form of the same utterance. Both are made of concepts. While the logical form of the utterance may be in need of complementation to be truth evaluable, the propositional form is not: It is the result of the enrichment of the logical form, a pragmatic process, which corresponds, to a minimum, to the attribution of referents, and to lexical disambiguation and can include enrichment processes, i.e., broadening and strengthening.

[7] I use small capitals, as is usual, to indicate concepts. The star after a concept indicates that it has undergone an operation of strengthening or broadening and has thus been replaced by the resulting ad hoc concept. Two stars after a concept (as in (11)) indicate that the ad hoc concept is the result of both broadening and strengthening operations, which have been applied in succession.

- They weakly communicate a wide array of implicatures (which is why they cannot be paraphrased without loss).
- The speaker used a metaphor because his thought was too complex to communicate literally.

It is not at all clear that these two features of the vintage account can be preserved in the current one, despite the fact that they have not been explicitly repudiated. Additionally, it is not clear that the notion of an ad hoc concept, as described by Carston (2002), is really coherent with the Relevance-Theoretic view of concepts, which has remained stable throughout the years, and which she claims to support.

29.5 Ad Hoc Concepts and Relevance Theory

From its origins, Relevance Theory has adopted a Fodorian (Fodor 1998) view of concepts according to which *concepts are atomic,* that is, they cannot be decomposed: They are *not* collections of features characteristic of the objects falling in the category corresponding to the concept[8]. In other words, *concepts are not definitions*. The standard Relevance-Theoretic view of concepts, supposedly based on Fodor's theory, claims that atomic concepts consist of an address or node in memory, giving access to three kinds of information:

- A *logical* entry, consisting of a set of inference rules (or meaning postulates), which capture analytic implications of the concept, usually not amounting to a definition.
- An *encyclopaedic* entry, consisting of a hodgepodge of assumptions and experiences.
- A *lexical* entry, specifying the phonetic, phonologic, and syntactic features of the linguistic encoding of the concept.

On the face of it, this may seem contradictory with the view that concepts are not collections of information. However, this is not necessarily correct. For instance, Millikan (2000), who also defends an atomic view of concepts (though on different grounds than Fodor, see Reboul 2007), makes a distinction between a *concept* and the information that can be gathered about the members of the corresponding category. She calls that set of information the *conception* and carefully distinguishes it from the concept. Arguably, the lexical entry is neither part of the concept (defined as a mental representation nomologically linked to the category), nor part of the conception: rather it has to do with the lexical item corresponding to the concept. But both the logical and the encyclopaedic entries fall in the conception and not under the concept itself, in keeping with Fodorian atomism.

As said above, Carston (2002) endorses this view of concepts. Yet, it is not clear that her account of the two pragmatic processes of broadening (also sometimes

[8] I shall use the term category for the extension of a given concept.

called *loosening*) and strengthening are entirely consistent with either the Fodorian view of concepts or the Relevance-Theoretic formulation of it, as can be seen from the following quotation (Carston 2002, p. 339):

> An *ad hoc* concept formed by strengthening a lexical concept seems to involve elevating an encyclopaedic property of the latter to a logical (or content-constitutive) status […]; an *ad hoc* concept formed by the loosening of a lexical concept seems to involve dropping one or more of the logical or defining properties of the latter.

On that description, the notion of an ad hoc concept seems to violate the Fodorian view of concepts, adopted by Relevance Theory, in what is central to it, i.e., its atomicity. That is, because Carston's definitions of strengthening and loosening heavily rely on the notion of "content-constitutive" or "defining" properties, in direct contradiction with Fodorian strictures. What is more, Carston claims that ad hoc concepts themselves are atomic, which, on her account of how they are built, seems plainly impossible.

Before I continue to dissect Carston's views on metaphor, I would like to make a (brief) digression on metaphor and paraphrase.

29.6 Metaphor and Paraphrase

As said above (Sect 29.2), it is a widely accepted fact that it is impossible to paraphrase a metaphor without loss. An intriguing question, however, is what is lost in the paraphrase relative to the metaphor. Another, easier, question is whether all metaphors are indeed impossible to paraphrase without loss. Let me first propose the following definition of a paraphrase (built from its dictionary definition in the French Larousse, 2011):

> A paraphrase is a different formulation of an utterance without any alteration of its meaning.

As pointed out above (Sect 29.2), one of the main criticisms of the vintage Relevance Theory account of metaphors against the classical accounts in terms of figurative meaning is that such accounts posit that there is a single legitimate meaning for a given metaphor, i.e., the figurative meaning. Thus, so-called figurative meanings are nothing more or less than paraphrases of the corresponding metaphors. In other words, such classical accounts are "paraphrastic" accounts of metaphors. But paraphrastic accounts are only tenable if paraphrasing a live metaphor, in agreement with the definition given above, does not actually change its meaning. There are, however, serious doubts that this is possible for all metaphors. For instance, it is most certainly not possible to paraphrase a live, creative, metaphor such as (1) without loss, which suggests that the meaning of the metaphor is altered in the paraphrase (for instance (2)). On the other hand, it is far from clear that this is the case for dead metaphors: thus, (4), (5), or (10) can be paraphrased as in (12), (13), and (14), respectively:

12. John does not care for other people's feelings.
13. John is a sweet and sensitive man, considerate of other people's feelings.
14. Caroline is a selfish and spoilt woman, unlikely to help.

I shall come back below to the distinction between live and dead metaphors. Right now, let me just note that the best way of destroying a live or creative metaphor is to paraphrase it, which raises a serious problem for any paraphrastic account of metaphors.

The central question, however, is why it is impossible to paraphrase a metaphor without loss. There are two (not incompatible) ways of answering it:

- The first answer, directly derived from vintage Relevance Theory, is that paraphrasing a metaphor leads to a loss of propositional effects.
- The second answer is that paraphrasing a metaphor destroys all the nonpropositional effects triggered by a metaphor.

I shall come back below (Sect. 29.8) to nonpropositional effects. Regarding propositional effects, the idea is that the paraphrase reduces a metaphor to a single interpretation (or implicature), while a metaphor, especially when it is creative, triggers a myriad of interpretations. Thus, paraphrasing a metaphor reduces this potential multiplicity to a single implicature, and, what is more, a *weakly* communicated implicature, hence, is an implicature whose recovery was not essential to the successful interpretation of the metaphor. (This is, of course, directly linked to the other specificity of metaphor in vintage Relevance Theory, that is, the impossibility of expressing literally the speaker's thought.) By contrast, nonfigurative utterances are usually possible to paraphrase because they do *strongly* implicate a single (or at most a few) proposition(s).

This was the vintage account, but in the current account, the *implicature*-based analysis has been replaced by an *explicature*-based analysis and thus the current account meets with a problem that seems very similar to that encountered by the figurative meaning account. Even though there is nothing to prevent a metaphor from having several explicatures, it does not seem likely that the explicature-based account allows as much liberty of interpretation to the hearer as the vintage Relevance-Theoretic account did. In other words, it is not clear why the explicature is not a paraphrase. Thus, it is not clear either that the current account can explain why paraphrasing a metaphor leads to an interpretive loss. Additionally, this problem is specific to the current account, being blocked in the vintage account by the notion of a weakly communicated implicature. Finally, the mechanisms of interpretation described in the current account seem rather more complicated than those of the vintage account.

Briefly, the two accounts share the idea that to interpret a metaphor (or any other utterance, given that both are continuist accounts) one should access the logical and encyclopaedic entries of the concept (e.g., PRINCESS) from the logical form of the utterance and select among the informations in the encyclopaedic entry the one appropriate given in the context (e.g., *A princess is lazy, spoilt, and unlikely to help*). However, in the vintage account, there was no modification of the conceptual entries themselves while in the current account, there is one (or two) modification(s) of the logical entry (leading to the ad hoc concept, e.g., PRINCESS**). Given that Relevance Theory is supposed to be geared toward cognitive economy, this is rather disturbing.

There is worse, however. The notion of an ad hoc concept seems to reintroduce some kind of mental paraphrase, on a par with theories of figurative meaning. It thus becomes difficult, not only to see what the notion of an ad hoc concept brings to the analysis, but also to explain why metaphors cannot be paraphrased without loss. How does CAROLINE IS A PRINCESS** differ from *Caroline is a spoilt and lazy woman unlikely to help* if PRINCESS** = spoilt and lazy woman unlikely to help? In addition, given that the notion of an ad hoc concept leads to the production of a *single* (true or false, but appropriate) explicature of a metaphor, it is difficult to see how this single interpretation is different from a figurative meaning. If this is the case, the current version of Relevance Theory only differs from the classical theory of figurative meaning in that it supposes that *all* utterances are interpreted through the production of ad hoc concepts.

29.7 Taking Stock

Thus, to sum up the conclusions reached up to now:

- The notion of an ad hoc concept is inconsistent with one of the basic tenets of Relevance Theory, i.e., the *atomicity* of concepts.
- If the explicature is a paraphrase of the metaphor (and it is hard to see how it could be anything else), then the current account contradicts one of the three main tenets of the vintage account, that is, the impossibility of paraphrasing metaphors (also widely recognized outside of Relevance Theory).
- Additionally, it also contradicts the vintage notion that metaphors are produced when there was no other way of expressing the speaker's thought.

There is more, however. There does not seem to be any way in the current account that metaphors can transmit an array of weak implicatures. This is because:

- On Relevance Theory, implicatures are produced as synthetic inferences from the context and the logical form of the utterance.
- The context is built (in part) from the encyclopaedic entries of the concepts in the logical form.
- On the current account, the logical form of the utterance is the explicature with the ad hoc concept.
- Ad hoc concepts, on Carston's account, have no encyclopaedic entry.

29.8 Nonpropositional Effects

Paraphrase and translation are two ways of the same saying:

15. Galileo said: "Eppur si muove."
16. Galileo said: "And yet it moves."
17. Galileo said: "And yet it is not static."

(16) is a translation and (17) a paraphrase of (15). Most utterances can be both paraphrased and translated without any major loss. However, as we have seen, metaphors cannot be paraphrased without loss. Nevertheless, they can be translated:

18. Romeo: But soft, what light through yonder window breaks?

It is the East, and Juliet is the sun.

19. Mais attends, quelle lueur passe par cette fenêtre?

C'est le levant et Juliette est le soleil.

Translation preserves the metaphorical effects of the metaphor while paraphrase destroys them. There are two possible explanations for that difference. The first goes through propositional effects and comes basically to the vintage Relevance-Theoretical account: The paraphrase reduces a metaphor to a single interpretation (or implicature) while a metaphor, especially when it is creative, triggers a myriad of interpretations. The second one goes through nonpropositional effects and says that the paraphrase destroys all the nonpropositional effects triggered by the metaphor. The notion that metaphors have nonpropositional effects is hardly new: Davidson (1978) already noted that metaphors have nonpropositional, sensory, effects. Davidson, indeed, thought that this was what metaphors were about and repudiated the idea that metaphors have any propositional effects. Guttenplan (2005) also gives sensory effects a prominent place in his account of metaphors.

Here, a potential problem with Sperber and Wilson's analysis, as well as with Carston's, lies in their choice of examples. Most of their examples are of dead metaphors and in the rare cases when they give examples of live metaphors, these are not analyzed. Dead metaphors have both few propositional effects and no nonpropositional effects. Live metaphors are quite different and are often accompanied by strong nonpropositional, sensory, effects, as can be seen by the following examples:

20. The sleep of reason begets monsters.

21. The yellow fog that rubs its back upon the window-panes
The yellow smoke that rubs its muzzle on the window-panes
Licked its tongue upon the corners of the evening
Lingered upon the peel that stand in drains
Let fall upon its back the soot that falls from chimneys
Slipped by the terrace, made a sudden leap,
And seeing that it was a soft October night,
Curled once about the house and fell asleep.
(T.S.Eliot, *The love song of J. Alfred Prufrock*)

(20) not only has visual effects, it was also accompanied by an illustration, appearing on the frontispiece of Goya's *Capricios*. (21), which is not, also gives rise to visual effects, and it is hard to read it without imagining a very big, very fat, yellow cat.

But such effects are not limited to metaphors, as can be seen from the following examples of haikus:

22. Cold winter shower
See all the people
Running

Across the Seta Bridge.
(Joso)

23. This snowy morning
That black crow
I hate so much…
But he's so beautiful!
(Basho)

Though some haikus are metaphorical, most are not, and the examples in (22) and (23) certainly are not and they both have strong sensory (visual) effects. Hence, metaphors are not the only utterances to have sensory effects. What is more, those sensory effects, given that nonmetaphorical haikus have them, are not specific to metaphors. What is, I will claim, specific to (live) metaphors, is the combination of an array of weakly communicated implicatures (in keeping with the vintage account) and sensory effects (as advocated by Davidson). Both of these features of metaphors do not lead to a discontinuist account: regarding the first one, the weak communication of a wide array of implicatures, Relevance Theory vintage account is perfectly adequate; regarding the second one, sensory effects, any account of it will have to explain why, *ceteris paribus,* (apparently) similar effects are to be found in nonmetaphorical utterances such as haikus.

For reason of space, this is not the place to try and give an account of why there are sensory effects in live metaphors and haikus (and presumably other types of utterances yet to be identified) and not in other utterances. My own rather tentative view is that this is due to the grounding of concrete concepts in perception (Reboul 2007). Recent work in neurolinguistics (e.g., Hauk et al. 2004; Nazir et al. 2008), has shown that the audition of different kinds of words (e.g., action words vs. object words) will activate the sensorimotor areas corresponding to the perception or action designated, which goes some way to corroborate this view. The differences noted above between paraphrase and translation regarding metaphors might thus be explained by the fact that changing the words of the sentence also changes the concepts in the logical/propositional form in the paraphrase, thus cutting access to the sensory effects associated to these concepts while changing the words in translation will not necessarily change the concepts, thus, leaving intact the sensory effects of the metaphor. What remains mysterious, however, is why these sensorimotor brain activations do not reach consciousness in the processing of most utterances, while they do in the case of, at least, metaphors and haikus, but that will be the matter for another chapter.

29.9 Conclusion

To sum up what I have tried to show in the present chapter, the vintage Relevance-Theoretic account of metaphor entirely failed to account (it did not even notice) the sensory effects, which seem typical of live metaphors. However, it did provide a rather convincing account of the propositional effects of metaphors, notably of live

metaphors, through the notion of an array of weakly communicated implicatures. This notion also allowed it to explain why paraphrastic accounts of metaphors, including classical accounts in terms of figurative meaning, will not work.

By contrast, the current account does not provide a convincing account of metaphors: It seems to fail to account for either the impossibility of paraphrasing metaphors, or the sensory effects of metaphors, and, indeed, it is hard to see why it is not merely a more or less cognitive or "up to date" version of figurative meaning accounts. Additionally, it seems more complicated than the vintage version in terms of the mechanisms involved, which makes it rather unattractive.

One central (though not specific) characteristic of metaphors is their visual effects. Both the vintage and the current Relevance-Theoretic accounts of metaphors entirely ignore it, possibly because, as do most accounts of metaphors, they tend to concentrate on the examples of dead metaphors, rather than live ones. So, an account of metaphor that will combine a healthy attention to their sensory effects without neglecting their propositional effects, relying on examples of live metaphors, is more necessary than ever.

Acknowledgments On December 21, 1990, I defended my Ph.D. thesis and was the first Ph.D. student to be supervised only by Kevin Mulligan in Geneva. My Ph.D. thesis was concerned with fiction and metaphor, and, though I came back to fiction fairly regularly throughout the years, it is only recently that my interest in metaphor was revived. So this chapter goes back to the year 1990 and to metaphor, and is, of course, dedicated to Kevin.

References

Barsalou LW (1983) Ad hoc categories. Mem Cognition 11(3):211–227

Carston R (2002) Thoughts and utterances: the pragmatics of explicit communication. Basil Blackwell, Oxford

Davidson D (1978) What metaphors mean. Crit Inquiry 5(1):31–47

Fodor JA (1998) Concepts: where cognitive science went wrong. Oxford University Press, Oxford

Goodman N (1970) Seven strictures on similarity. In: Foster L, Swanwon JW (eds) Experience and theory. Duckworth, London, pp 19–29

Guttenplan S (2005) Objects of metaphor. Clarendon, Oxford

Hauk O, Johnsrude I, Pulvermüller F (2004) Somatotopic representation of action words in human motor and premotor cortex. Neuron 41:301–307

Millikan RG (2000) On clear and confused ideas: an essay about substance concepts. Cambridge University Press, Cambridge

Nazir TA, Hauk O, Jeannerod M (2008) The role of sensory motor systems for language understanding. J physiology-Paris 102:1–3

Reboul A (2007) Langage et cognition. Presses Universitaires de Grenoble, Grenoble

Searle J (1985) Metaphor. In: Searle J (ed) Expression and meaning: studies in the theory of speech acts. Cambridge University Press, Cambridge, pp 76–116

Sperber D, Wilson D (1995) Relevance: communication and cognition, 2nd edn. Basil Blackwell, Oxford

Chapter 30
Syntactic Cartography and the Syntacticisation of Scope-Discourse Semantics

Luigi Rizzi

Abstract Cartographic studies aim at drawing maps as precise and complete as possible of syntactic configurations. Such maps interact with principles of syntactic computations, and provide crucial information for the interfaces with sound and meaning. In this chapter, I will concentrate on the role that cartographic representations have in the expression of interpretive properties and, in particular, in the organization of informational structures for the proper articulation of discourse and dialogues, and in the assignment of scope to operators. I will then illustrate the criterial approach to scope-discourse semantics, an approach which implies fully transparent syntactically generated interfaces with interpretive systems, and thus is sometimes said to "syntacticize" the expression of such interpretive properties. I will then compare this approach to possible alternatives which involve more opaque interfaces and assume more complex computations in postsyntactic interpretive systems.

Keywords Syntactic cartography · Merge · Syntactic movement · Scope-discourse semantics

30.1 Introduction

The cartography of syntactic structures is a program which aims at drawing maps as precise and complete as possible of syntactic configurations. Cartographic studies over more than a decade have brought to light the complexity of the structure of sentences and phrases, but also the simplicity of the underlying generative mechanisms: Complex representations arise from the recursive application of a very elementary combinatorial procedure ("merge" in Minimalist terminology: Chomsky 1995) operating on the substantive lexicon (nouns, verbs, adjectives, …) and on a very rich functional lexicon (Belletti 2004, 2009; Cinque 1999, 2002; Rizzi 1997,

L. Rizzi (✉)
University of Siena, Siena Italy
e-mail: rizzil@unisi.it

University of Geneva, Geneva, Switzerland

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8_30,

2004a, b, and, for recent assessments of the cartographic projects, Cinque and Rizzi 2010; Shlonsky 2010).

While technical details of cartographic representations may look remote from the philosopher's preoccupations, the interface between syntax and interpretation, and, more generally, the systematic relation between form and meaning, is not (Mulligan 1987, 2006). So, in this chapter, I will concentrate on the role that cartographic representations have in the expression of interpretive properties and, in particular, in the organization of informational structures for the proper articulation of discourse and dialogues, and in the assignment of scope to operators.

I will first illustrate the duality of semantic properties that natural languages express, and the division of labor between the fundamental computational mechanisms used for their expression. I will then illustrate the criterial approach to scope-discourse semantics, an approach which implies fully transparent syntactically generated interfaces with interpretive systems, and thus is sometimes said to "syntacticize" the expression of such interpretive properties. I will then compare this approach to possible alternatives which involve more opaque interfaces and assume more complex computations in postsyntactic interpretive systems.

30.2 Two Types of Semantic Properties and Their Expression in Syntactic Representations

The interpretation of natural language expressions revolves around two broad kinds of interpretive properties:

1. Properties of argumental semantics: who does what to whom in the event referred to by a sentence, what thematic roles are expressed, etc.
2. Properties of scope-discourse semantics: the scope of operators and the expression of discourse-related properties linked to the informational organization of the sentence, such as topicality and focus.

To illustrate the first kind, consider the following sentences:

(1) a. It was raining
b. John left
c. John saw Mary
d. John gave a book to Mary

Each verb expresses a kind of event which can be depicted as a little scene involving a certain number of participants: 0 in (1)a (the subject pronoun is not referential here, it is just a place holder to satisfy the formal requirement that all sentences must have a subject: Rizzi 2006, Rizzi and Shlonsky 2007 and much related literature), 1 in (1)b, 2 in (1)c, 3 in (1)d. This is the argument structure of the verb, expressing who does what to whom. The roles of participants can be further differentiated by certain qualitative labels, the thematic roles: So John is the agent in (1)d, the participant who causes the event acting according to a conscious plan, and the experiencer

in (1)c, the participant who undergoes a certain perceptive of psychological experience; Mary is the patient in (1)c (or, in other terminologies, the theme) and the goal in (1)d, etc.

Consider now the following sentences, with (2)b and c, obtained from (2)a through some formal manipulations that I will go back to:

(2) a. John gave your book to Mary
b. Your book, John gave ___ to Mary
c. It is your book that John gave ___ to Mary.

These sentences share the same argument structure: There is an event of giving involving three participants, John, your book, and Mary. In particular, the phrase *your book* has the same argumental status; it is the patient of *give* in the three cases. But the very same expression has different informational properties in the three cases. Such properties can be highlighted by creating little discourse contexts which enforce a particular organization of the information that is exchanged by the participants in the dialogue. *Your book* is naturally interpreted as part of the new information expressed by the predicate in (2)a, i.e., (2)a could be appropriately used to answer a question like (3):

(3) Q: What did John do? (And Bill?)
A: John gave your book to Mary (as for Bill, I don't know what he did)

(The contrast John/Bill is introduced here to make fully natural the reiteration of *John* as the subject of (2)a: If there was no such contrast, the natural choice would be to use a pronominal subject: *He gave your book to Mary*; if a contrast is present, it is natural to reiterate the proper name as a kind of contrastive topic.)

In (2)b, *your book* is interpreted as the topic of the sentence, taking up and making salient a referent already given in the discourse. So, (2)b could appropriately answer a question like (4), which introduces a certain book in the discourse context (again, the contrast is introduced to make the overt expression of the topic in the answer nonredundant):

(4) Q What did John do with my book? (and with Bill's?)
A: Your book, John gave ___ to Mary (as for Bill's book, I don't know)

A natural interpretation of the cleft construction in (2)c is that *your book* is the contrastive focus, correcting an assumption that the speaker imputes to the hearer, for instance because the latter just expressed it (on the syntax and interpretation of clefts see Belletti 2008 and references quoted there). Consider for instance the following dialogue between speakers A and B:

(5) A: I know that John gave Peter's book to Mary…
B: (no,) it's your book that John gave to Mary (, not Peter's)

Here speaker B corrects speaker A by uttering (2)c, possibly making the contrast explicit through the negative tag.

So, the same phrase, always holding the same argumental role, can assume very different informational roles and function differently in discourse depending on the position in which it is pronounced.

In addition to expressing distinct informational properties, the patient of *give* can acquire operator-like status in other constructions in which it does not appear in its canonical thematic position after the verb, but is dislocated to the front in relatives or interrogatives (or other left peripheral constructions):

(6) a. The book that John gave ___ to Mary (is very nice)
b. Which book did John give ___ to Mary?

These sentences will have logical forms roughly like the following:

(7) a. The unique entity x, x a book, such that John gave x to Mary (is very nice)
b. For which x, x a book, John gave x to Mary?

(where uniqueness in (6)a is of course restricted to the relevant books in the discourse context).

In conclusion: An element bearing a particular role of argumental semantics can assume different discourse or operator functions which typically correspond to particular positions in the clausal structure. These are the properties that Chomsky (2000, 2004) calls "properties of scope-discourse semantics," and which I will refer to later on with the technical term "criterial properties." How can the properties of argumental and scope-discourse (criterial) semantics be expressed formally?

30.3 Merge and Move

Words are strung together to form phrases which are hierarchically organized. So, for instance, in:

(8) John will give a book to Mary

[*A book*] and [*to Mary*] form phrases which can be manipulated as units (for instance, can be focused in the cleft construction: *It is [a book] that John will give ___ to Mary, It is [to Mary] that John will give a book* ___) while *give a* and *to Mary* do not form phrasal units. There must be an algorithm building the hierarchical structure of the sentence, an algorithm endowed with recursive properties (because we can indefinitely expand a sentence: (8) can be part of a larger sentence like *Peter thinks that John will give a book to Mary,* etc.). A number of rather different recursive structure building algorithms have been considered in the history of generative grammar: generalized transformations, phrase structure rules, *X*-bar schemata,…. The minimalist program (Chomsky 1995) has come to the conclusion that the structure building algorithm is the simplest combinatorial rule one can imagine:

(9) Merge: take two elements A and B and string them together to form the phrase [A B]

Merge can take two elements from the lexicon, say *hit, Bill,* to form the verb phrase *[hit Bill]*. It can recursively reapply to string together the structure just formed with another element taken from the lexicon, e.g., *will,* to form the phrase *[will [hit Bill]]*, and then reapply again to form *[John [will [hit Bill]]]*, and so on.

Merge is intimately related to the expression of argument structures and the assignment of thematic roles. So, a verb like *hit* has two roles to assign, agent and patient. When *hit* is merged with a nominal expression like *Bill* to form the verb phrase *[hit Bill]*, the patient role is discharged to *Bill*. The structure thus created can be further merged with another nominal expression, *John,* forming the expression [*John [hit Bill]*]; here, *John* receives the remaining role of an agent. We can think that all the assignment of thematic roles works like that: Merge creates the local configurations between assigners and assignees for the expression of argumental semantic properties. So, a head assigning thematic roles (typically a verb, but in fact any lexical item can be an assigner) assigns the roles specified in its lexical representation to its immediate dependents, in the local configurations created by repeated applications of merge. So, thematic assignment is strictly local, with the relevant local configurations provided by merge.

Consider now the assignment of scope-discourse semantic properties. One could think that such properties as topic–focus, etc. are superimposed to the hierarchical structures created by merge. This may indeed happen in some cases (e.g., if an element can be focalized in situ, without being displaced from its argument position), but this is by no means the typical procedure. What typically happens is what we have already seen in (2)b and c, and also in (6)a and b: The element is displaced from its thematic position to another position, typically in the initial periphery of the sentence, where it receives its appropriate scope-discourse property: topic, contrastive focus, relative or interrogative operator.

So, displacement, or movement, is systematically used by natural languages for assigning the two kinds of interpretive properties to an element. The element is merged in a position in which it receives its argumental status, a thematic role; then it is moved to another position dedicated to a particular scope-discourse property.

What is movement? The traditional view directly implements the metaphor of physical displacement: The element is taken from its original position, which is vacated, and moved to a higher position in the syntactic tree. But there are other views of movement which are less faithful to the physical metaphor. For instance, in Minimalism "movement" involves (1) the identification of a candidate (through a search operation, which I will go back to later on) and (2) an application of merge remerging the identified candidate with the structure created so far. The position identified as a candidate is not physically displaced: It continues to host a complete but silent (unpronounced) copy of the remerged phrase. For instance, the derivation of a sentence containing a topicalized phrase like (2)b is as follows. Starting from a structure like:

(10) [John gave [your book] to Mary]

Search identifies the phrase to be topicalized *your book*. Then the phrase is merged with the whole structure and the original position remains filled by an unpronounced copy of the phrase (notated through the angled brackets in (11)):

(11) [your book] [John gave <your book> to Mary]

Representation (11) expresses well the way in which the language assigns the dual semantic properties to the phrase *your book:* It occurs twice, once in the argumental

position and once in the scope-discourse position, and thus it picks up the thematic role "patient" and the status of topic.

In this view, movement reduces at least in part to merge: The common terminological practice distinguishes between "external Merge" (9), and "internal merge," which still involves two elements A and B which are strung together; the difference with external merge is that here the two elements are not external to each other before the operation, but rather one of them is internal to the other. So, we could depict the operation of internal merge as follows:

(12) $$[_B \ldots A \ldots] \rightarrow [\, A [_B \ldots <A> \ldots \,]],$$

where A is first selected by search within B as a candidate for movement, then A is remerged with the whole structure B, with the original occurrence of A becoming the silent copy <A>. In fact, merge performs exactly the same operation in (9) and (12), creating the structure [A B]. So, it would perhaps be more appropriate to say that there is a fully unified operation merge, but the search operation identifying the candidates for the application of merge can be external (looking at elements that are external to each other, for instance two lexical items) or internal (looking at an element internal to the other element, as in (12)). In any event, we will continue to use the standard terminology, keeping in mind that the unification of movement and structure building under merge may be even more complete than the terminology suggests.

30.4 The Criterial Approach to Scope-Discourse Semantics

We can now come to the key issue. What does it mean that a position is "dedicated" to a certain kind of interpretive property? In the case of argumental semantics, things are rather uncontroversial: Thematic assignment is a matter of local head-dependent relation, a verb assigns its thematic roles to its immediate dependents (its specifier and complement, in the traditional terminology of the *X*-bar notation). So, the structure created by merge expresses immediately and transparently that in [*John [hit Bill]*] John is the "hitter" and Bill is the "hittee," or the agent and patient, respectively (some current systems use explicit thematic labels like agent and patient, others do not, but I do not see more than a notational decision in this choice: Any system must express the notional content of agent, patient, etc. somewhere in the system, whether or not explicit thematic labels are used).

More controversial is the syntactic expression of scope-discourse semantic properties. Consider the following constructions expressing distinct scope-discourse properties:

(13) a. Your book, John will give ___ to Mary
b. It is your book that John will give ___ to Mary, not Peter's book
c. The book that John will give ___ to Mary (is interesting)
d. Which book will John give ___ to Mary?
e. What a nice book John gave ___ to Mary!

How does the nominal expression including the lexical specification *book* receive its interpretation of topic, contrastive focus, relative operator, interrogative operator, exclamative operator, respectively?

Here, I will present the criterial approach, a view that has been largely assumed and supported by cartographic studies, and will then briefly compare this approach with possible alternatives. According to the criterial view, the assignment of scope-discourse properties is done on a strictly structural basis, much as the assignment of argumental properties. The approach assumes a set of functional heads which populate the initial periphery of the clause. Such heads, Top(ic), Foc(us), Rel(ative), Q(uestion), Exc(lamative) have a dual function, internal to syntax and relevant for the interfaces with sound and meaning:

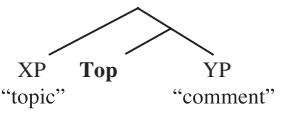
1. They attract a phrase to their specifier (in terms of the Minimalist approach to movement, they activate a search of a candidate phrase, which then undergoes internal merge).
2. They trigger specific interpretive routines at the interfaces, determining the interpretation on the meaning side, as well as the assignment of the special, marked intonational contours which make such constructions easily detectable for the hearer.

In terms of the syntactic representations involved, the approach claims that sentences such as (13) have representations such as the following (in which the copy theory of traces is adopted; NB: *that* is separated from Foc and Rel for clarity in (14)b and c, but it may very well be a particular morphological realization of the criterial heads here; analogously, Q in (14)d may well be the position targeted by the inverted auxiliary *will*):

(14) a. [Your book] **Top** [John will give <your book> to Mary]
b. It is [your book] **Foc** [that John will give <your book> to Mary], not Peter's book
c. [The book] **Rel** [that John will give <the book> to Mary] (is interesting)
d. [Which book] **Q** [will John give <which book> to Mary]?
e. [What a nice book] **Excl** [John gave <what a nice book> to Mary]!

The crucial elements of such structures are the functional heads Top, etc., which mediate between the two constituents isolated by the brackets. Such heads have the syntactic function of attracting the topic, the focus, etc. in clause initial position, and carry interpretive routines used on the other side of the interface. For instance, Top goes with the following:

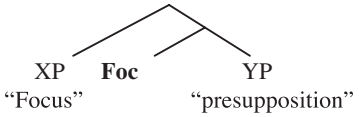
(15)



(15) expresses the topic–comment articulation: The phrase attracted to the specifier position of Top is interpreted as the topic, designating a referent chosen between the entities familiar from the previous context, made salient, and about which something is said, i.e., a comment is made. YP, the complement of Top, expresses the comment, which typically contains the new information (more fine-grained analyses on the typology of topic positions and their interpretations are offered in Benincà and Poletto 2004; Bianchi and Frascarelli 2009; Frascarelli and Hinterholzl 2007).

Analogously, the Foc head triggers an interpretive routine which expresses the contrastive focus–presupposition articulation:

(16)



So, in (14)b *your book* is interpreted as (contrastive) focal information, an information which the speaker assumes to be new, and also somehow falling outside the expectations of the hearer (which can be explicitly expressed and corrected by the negative tag); this against the background of presupposed, shared knowledge. In other words, when I utter (14)b, I assume that my interlocutor shares the knowledge that John will give something to Mary, and I assert that this something, contrary to my interlocutor's expectations, is your book (the exact interpretation of left peripheral focus positions is parametrized in part: e.g., Cruschina (2008) shows that the Sicilian dialect uses the left peripheral position to express simple new information focus).

The parallel with the assignment of argumental semantic properties of this "structuralist" approach to scope-discourse semantics should be clear: Top attributes the status of topic and comment to its specifier and complement much as *hit* attributes agent and patient to its dependents. The parallel holds with some systematic differences separating the two cases:

1. Thematic assignment is typically done by a lexical head (primarily, a verb) while scope-discourse assignment is done by a functional head.
2. The relevant syntactic configuration is created by external merge for thematic assignment, and by internal merge (movement) for scope-discourse assignment.

So, there is a systematic division of labor between the functional and substantive lexicon, and between external and internal merge in expressing the duality of semantic properties that characterize the interpretation of natural language expressions.

What arguments can be given in favor of such a structuralist view of assignment of scope-discourse semantic properties? A straightforward argument is offered by the fact that in some languages the system of functional heads assumed here is expressed by overt morphemes. Consider the following cases:

(17) a. Ik weet niet [wie *of* [Jan ___ gezien heeft]] (Dutch varieties, Haegeman 1996)
'I know not who Q Jan seen has'
b. Un sè [do [dan lo *yà* [Kofi hu ì]]] (Gungbe, Aboh 2004)
'I heard that snake the Top Kofi killed it'
c. Un sè [do [dan lo *wè* [Kofi hu ___]]] (Gungbe, Aboh 2004)
'I heard that snake the Foc Kofi killed '
d. Der Mantl [den *wo* [dea Hons ___ gfundn hot]] (Bavarian, Bayer 1984)
'The coat which R the Hans found has'
e. Che bel libro *che* [ho letto ___]! (Italian)
'What a nice book Excl I read'

In many dialectal varieties of Dutch, wh elements in embedded questions can co-occur with the question marker *of* (if), as in (17)a, and also with the complex form *of* + *dat*. While *dat* is not specific to questions (e.g., it can also introduce declaratives), *of* is specific, so, it appears to be a good candidate for an overt realisation of the Q head attracting the wh element to its Spec. The examples, (17)b and c from the African language Gungbe illustrate the case, quite frequent crosslinguistically, of a language using overt topic and focus markers (*yà* and *wè,* respectively), good candidates for the overt expression of the Top and Foc heads of (15), (16). The Bavarian dialect illustrates a property rather commonly found in dialectal varieties of German, with the locative wh elements *wo* specialized to introduce relatives; in this variety *wo* can also co-occur with the relative pronoun *den,* which sits in its Spec.

Standard Italian generally disallows co-occurrence of a preposed operator and a complementizer particle, except in exclamatives, as in (17)e. The particle here is *che,* a complementizer morpheme which clearly is not specialized for exclamatives (as it can also introduce simple declaratives). Still the possible co-occurrence with the operator is specific to exclamatives in the standard variety. So, questions and exclamatives are disambiguated by the presence or absence of this element (and of course, the intonation pattern is very different in the two cases, both on the operator and in the clause defining its scope domain):

(18) a. Che macchina hai comprato? (only question)
'What car did you buy?'
b. Che macchina *che* hai comprato! (only exclamative)
'What (a) car you bought!'

In view of such contrasts as (18)a and b, this particular occurrence of *che* may be seen as a particular lexicalization of the Excl head, attracting the exclamative operator to its Spec, and contrasting in this respect with the Q head, null in Italian (alternatively, *che* may lexicalize a lower position, possibly Fin in the system of Rizzi 1997, but its presence in the C system is contingent upon the presence of a higher Excl head, a contingency expressible, e.g., through a kind of agreement—or search operation across heads in the C system).

In English, the difference between questions and exclamatives is also signaled, although more indirectly, as the Q particle attracts the auxiliary and inversion occurs (in main questions) while the Excl particle does not trigger inversion (*What car did you buy*? vs. *What* (*a*) *car you bought*!).

In sum, the widespread existence of cases like (17) across languages provides very simple evidence for representations like (14), (15), (16), etc., as in (17) such heads as Top, Foc, Q, Rel, Excl are overtly expressed. We may thus assume that in

languages like English the interface with scope-discourse semantics also involves representations of this kind, except that the relevant heads are left unpronounced, a familiar and widespread kind of low-level parametric option (for instance, proper names occur without determiner in standard Italian or standard German, but appear with the determiner in many local varieties: *la Maria, der Hans,* etc.: again, a low-level parametrisation on the overt/non-overt character of a functional item seems to be involved, and such cases are innumerable).

Why are such heads, and the relevant configurations, called "criterial"? The term is an extension of Chomsky's (1981) Theta-criterion for argumental semantics, and refers to the specifier-head configuration which must be created with such constructions, e.g., the structures in (15) and (16): In a sense, the criterial heads act as "scope markers" for the phrases they attract to their Specs. The relevant configuration defines a "criterion," The Q criterion (in fact, originally called Wh criterion) was initially proposed some 20 years ago (Rizzi 1991; see also Rizzi 2000), and then the approach was extended to the whole family of left peripheral movements (focus criterion, topic criterion, etc.). Aboh (2010) has rephrased the criterial approach in terms of the minimalist program. We thus have criterial features (Q, R, Top, Foc, Excl) which act as attractors of phrases endowed with matching features, and trigger certain interpretive routines on both interfaces:

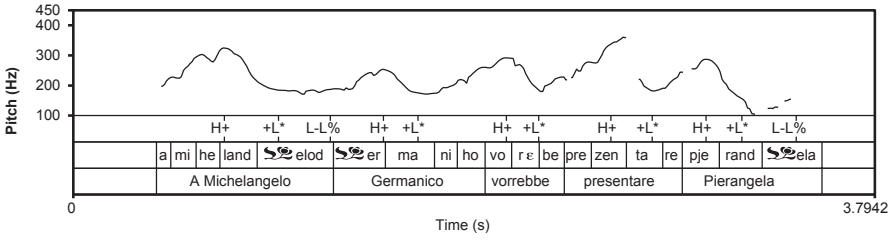
(19)a X_{CritF} is part of the numeration, triggers an internal search for XP_{CritF}; the XP_{CritF} thus identified undergoes internal merge to the Spec of X_{CritF}, for CritF = Q, R, Top, Foc, Excl,….

b X_{CritF} carries explicit instructions concerning how its dependents (Spec and complement) must be interpreted by the interface systems dealing with sound and meaning.

(Rizzi 1991; Aboh 2010)

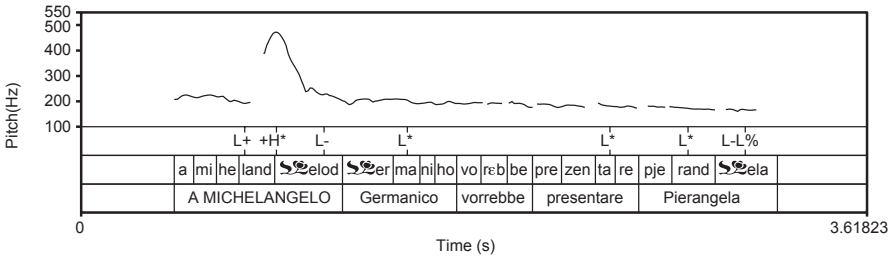
At the interface with semantics and pragmatics, the routines take the shape of (15) and (16), indicating which phrase must be interpreted as topic, comment, etc. There are also significant effects on the interface with sound, as criterial configurations are often highlighted by very salient special prosodic contours, easily detectable from the signal. Bocci (2009; Frascarelli 2000). has conducted a theoretical and experimental study on such contours in Italian (See also Frascarelli 2000). The typical contours associated with topic–comment and (contrastive) focus–presupposition in Italian are the following:

(20) Pitch contour of "Topic–Comment" in Italian (from Bocci 2009)



A Michelangelo (Top), Germanico vorrebbe presentare Pierangela
'To Michelangelo (Top), Germanico would want to introduce Pierangela'

(21) Pitch contour of "Focus—Presupposition" (from Bocci 2009)

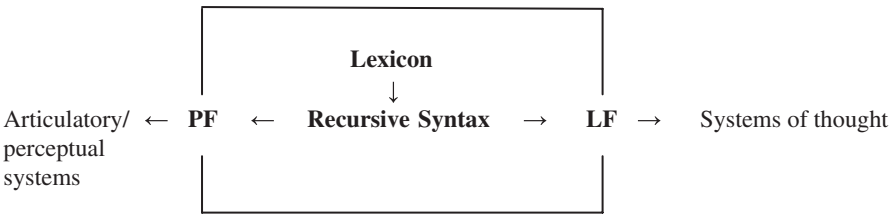


A MICHELANGELO (Foc) Germanico vorrebbe presentare Pierangela (, non a Piero)
'TO MICHELANGELO (Foc) Germanico would want to introduce Pierangela (, not to Piero)

The topic is marked by a certain prosodic prominence, followed by a "hilly" contour of the comment, as in (20). The contrastive focus is marked by a more pronounced prominence, followed by the complete flattening of the contour of the presupposition, as in (21). Bocci (2009) determined such contours experimentally, and proposed a system of prosodic rules at the phonological interface which "read" the structures passed on from the syntax and, capitalizing on the criterial heads and features, proceed to assign the appropriate contours.

The criterial view has been characterized as the attempt to "syntacticise" as much as possible the aspects of scope-discourse semantics, in that fundamental scope-discourse interpretive properties are traced back to basic syntactic configurations in a transparent and straightforward manner. Syntax wears interface properties on its sleeves, as it were. In fact, not only scope-discourse semantics, but also the prosodic properties are transparently read off from syntactic representations in this approach, as we have just seen. One important characteristic of this system is that the two interfaces are solely connected by syntax, any other connecting device directly relating, say, intonation and pragmatics can be dispensed with. That is, no other connecting line is required on top of the minimal connections expressed by the following classical articulation:

(22)



The box of linguistic computations includes a lexicon (divided into two components: contentive and functional) and recursive syntax. The system computes representations of phonetic form (PF) and logical forms (LF), to be understood as partial representations of sound and meaning inasmuch as such properties are grammatically determined. Such representations are further elaborated by other (language independent) systems on both sound and meaning sides, which use grammar-determined representations for communication, socialization, the expression of thought, play, art, and whatever use humans make of their linguistic abilities. In the approach I have presented, sound and meaning are solely mediated by syntax, also as far as scope-discourse properties are concerned. The uniqueness of the syntactic connection is uncontroversially assumed for argumental semantics: Nobody questions the fact that, say, the subject of a passive sentence (*John was hit by Bill*), pronounced at the beginning of the clause, is interpreted as the patient through the mediation of syntax (movement to a position distinct from the thematic position), and no syntax-independent link between PF and LF must be established to express this property. This mediating role of syntax applies fully to scope-discourse semantics as well, under the criterial approach.

The "syntacticization" effect is immediately visible by looking at diagram (22): Under the criterial approach, very little computation is needed after the LF interface to establish such properties as topicality and focus, as the properties are already transparently expressed by syntactic representations in the format of (15), (16), etc. This economy of postsyntactic computation is a hallmark of the criterial approach: Other systems would inevitably require a more structured set of postsyntactic interpretive rules to ensure the proper interpretation of syntactic configurations which, in such alternative systems, would be (more) opaque in the syntactic expression of scope-discourse interpretative properties.

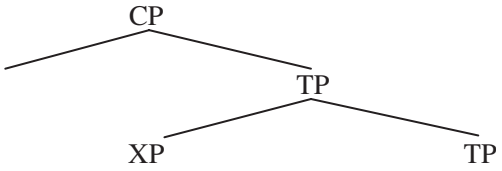
30.5 An Alternative

In conclusion, let us focus for a moment on such alternative systems, and see what their consequences are. A non-cartographic, non-criterial approach would try to get away with more impoverished syntactic representations. For instance, instead of postulating a structured left periphery of the clause, populated by a system of well-differentiated functional heads, one could go back to the traditional assumption that topicalisation, focalization, and simple preposing of an adverbial (neither properly topical nor focal: see below) have essentially the same syntactic representation, for instance involving the adjunction of the preposed element to the clausal category (labeled TP here, as in much minimalist practice):

(23) a [$_{TP}$ Your book, [$_{TP}$ I will read ___ tomorrow]]
b [$_{TP}$ YOUR BOOK [$_{TP}$ I will read ___ tomorrow]] (, not Peter's)
c [$_{TP}$ Tomorrow [$_{TP}$ I will read your book]]

That is, the three cases would share the same syntactic configuration in this syntactically impoverished approach, with the preposed element XP attached to the TP in an adjunction structure under the C(omplementizer) system:

(24)



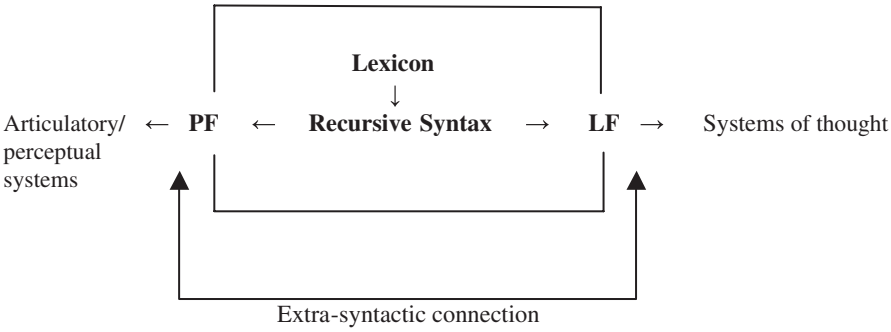
Cartographic analyses, on the other hand, assign three distinct representations here, geometrically analogous (they are all merge generated) but distinct in terms of the nature of the head defining the construction (in cartographic representations there will also be additional "space" between the criterial layer and the TP, defined by Fin, the head delimiting the lower bound of the C system; I omit this detail for simplicity here):

(25) a [$_{\text{TopP}}$ Your book **Top** [$_{\text{TP}}$ I will read ___ tomorrow]]
b [$_{\text{FocP}}$ YOUR BOOK **Foc** [TP I will read ___ tomorrow]] (, not Peter's)
c [$_{\text{ModP}}$ Next week **Mod** [$_{\text{TP}}$ I will read your book]]

Top and Foc have already been introduced; Mod(ification) in (25)c is the head which defines the position which a highlighted adverbial is moved to, a position which is neither properly topical (because the discourse context may very well lack any reference to particular weeks to ground proper topicality of the temporal adverbial) nor contrastively focal (because here *next week* is not necessarily contrasted to any particular belief that the speaker may impute to the hearer). If (25) is adopted, the interpretive properties are transparently expressed by the syntactic representation, which also directly guides the assignment of the appropriate prosodic contour through a Bocci-style rule system.

If the non-cartographic representations (23) are adopted, much of the interpretive work must be done postsyntactically: Rules must be postulated which assign to uniform representations like (23) (sharing configuration (24)) the interpretation of topic–comment, focus–presupposition and adverbial highlighting. Moreover, as distinct prosodies must be assigned in these cases (as is particularly clear in the case of focus), special direct connections must be postulated between semantics-pragmatics and phonetics to ensure the proper contour assignment: In this system syntax is too impoverished to play this connecting role, as the uniform structure (24) is assigned to (23)a, b, c, so if the post-PF component assigns a focus intonation to the preposed phrase, this information must be transferred to the post-LF interpretive component, hence an extra-syntactic connecting line must be assumed by such systems (it does not matter if the connection is from PF to LF, or vice versa):

(26) A system with impoverished, uniform syntactic representations:



Clearly, on grounds of simplicity, the organization in (22) is preferable, all other things being equal. There are also more specific arguments in favor of cartographic representations and the criterial approach. In certain contexts, the three constructions of (23) have very different syntactic properties, a fact that is unexpected given the structural uniformity postulated by the "impoverished syntax" approach. Consider for instance the opposite consequences that topicalisation and adverb preposing have with respect to so-called anti-adjacency effects. The following (27)a illustrates a *that*-trace effect, the fact that subject extraction is barred across an overt complementizer like *that* in standard varieties of English. As Bresnan (1977) observed, a preposed adverbial improves the acceptability considerably, as in (27) b, while an embedded topicalisation does not have a positive effect on acceptability, as in (27)c (on anti-adjacency effects, also called "adverb effects," see Rizzi 1997 and references quoted there):

(27) a * This is the man who I think that ___ will sell his house next year
b This is the man who I think that, next year, ___ will sell his house
c * This is the man who I think that, his house, ___ will sell this year

If the representations of topicalisation and adverb preposing are structurally different, as in (25), one may capitalize on the difference to capture the opposite consequences the two constructions have in alleviating *that*-trace effects (this is the line of analysis proposed in Rizzi 2009, capitalizing on the devices for subject extraction analyzed in Rizzi and Shlonsky 2007). If the syntactic representations are indistinguishable, as in (23) and (24), then syntax does not offer any straightforward basis for capturing the contrast in (27). We thus have purely syntactic reasons for choosing the well-differentiated representations assumed by cartographic studies and the criterial approach.

In this brief comparison, I have contrasted maximally different approaches like the one assuming a full-fledged cartographic representation and one assuming fully indistinguishable representations for different kinds of preposings. Comparing radi-

cally different approaches is useful because it allows us to better clarify the issues. But of course, various intermediate cases can be imagined. For instance, one could entertain an approach whereby the structural configuration is exactly the same, say an adjunction structure like (24), but the different cases of preposing are partly differentiated through the involvement of different morphosyntactic features: For instance, the T head could be optionally endowed with the features of topicality, focus, or modification, thus triggering the relevant adjunction to the TP projection. One could then try to appeal to such featural differences to express the observed differences in syntactic behavior, e.g., with respect to the anti-adjacency effect.

Putting aside the question of whether such differences could be naturally expressible in this way, fundamental empirical differences remains in the tree geometry between a cartography-based approach and an approach exploiting featural differences in otherwise uniform representations like (24): The cartographic approach based on criteria predicts a systematic biuniqueness between heads and specifiers while the alternative approach allows for the possibility of a single non-dedicated head (say T), supporting several specifiers/adjoined positions with different scope-discourse functions. Again, languages with overt topic–focus markers like Gungbe support the more restrictive approach. For instance, topic and focus can co-occur in Gungbe, but in a strict order and with each element supported by the relevant marker:

(28) … dò Kòfí *yà* gànkpá mè *wè* kpònòn lé sú – ì dó
'…that Kofi Top PRISON IN Foc policemen Pl shut him there' (Gungbe: Aboh 2004)

This is immediately expected under the cartographic approach while the occurrence of such elements as yà and wè would require auxiliary hypotheses under a structurally uniform approach based on TP adjunctions (how the strict ordering of projections arises is a separate issue: see Abels 2010; Haegeman 2011 for approaches involving locality principles, and Cinque and Rizzi 2010 for discussion).

30.6 Conclusion

Natural languages typically assign to expressions two kinds of interpretive properties: properties of argumental semantics and properties of scope-discourse semantics. This duality of semantic properties is reflected by dedicated mechanisms in the lexical organization (functional and contentive lexicon) and in the articulation of syntactic computations (external and internal merge).

Cartographic research has put on focus, at the same time, the elementary mechanisms appealed to by natural languages for the expressions of such properties, and the global syntactic configurations which arise from such mechanisms. What has emerged is, on the one hand, the richness of syntactic representations, which can only be captured by very detailed and refined structural maps; and, on the other hand, the very simple nature of the generating mechanisms, exploiting merge and only requiring very simple specifications on the functional heads. Given the complex nature of the task of expressing both types of properties, natural languages

thus seem to have selected mechanisms favoring local simplicity (very elementary specifier–head configurations agreeing on simple featural specifications), at the price of tolerating significant global complexity, generated by the reiteration of the computational atoms: Natural languages systematically exploit movement (internal merge), which allows an element to occur in different positions, picking up elementary interpretive properties in distinct dedicated positions. In this chapter, I have looked at mechanisms responsible for the expression of scope-discourse properties, expressed in terms of the criterial approach, which has turned out to be particularly congenial to cartographic research. Scope-discourse properties are directly expressed by structural configurations, much as argumental properties are: There is a set of dedicated functional heads (Top, Foc, Mod, Q, Rel, Excl,…) which undergo merge with specifiers and complements, and trigger transparent interpretive routines at the interfaces determining the interpretations of their structural dependents as topic, comment, focus, presupposition, operator of a given kind and its scope domain, etc. This view is sometimes said to involve the "syntacticisation" of scope-discourse semantics, in that it involves the creation, in the syntax, of fully transparent syntactic interfaces, requiring only straightforward additional computation in the interpretive systems. I have discussed some conceptual and empirical properties of this approach which favor it, in my opinion, in comparison to imaginable alternatives.

References

Abels K (2010) The Italian periphery: a view from locality. Ms, UCL, London
Aboh E (2004) The morphosyntax of complement-head sequences. Oxford University Press, New York
Aboh E (2010) Information structuring begins with numeration. Iberia 2(1):12–42
Bayer J (1984) Comp in Bavarian. Linguistic Rev 3:209–274
Belletti A (ed) (2004) Structures and beyond: the cartography of syntactic structures, vol 3. Oxford University Press, New York
Belletti A (2008) The CP of clefts. Rivista di Grammatica Generativa 33:191–204
Belletti A (2009) Structures and strategies. Routledge, London
Benincà P, Poletto C (2004) Topic, focus and V2: defining the CP sublayers. In: Rizzi L (ed) The structure of CP and IP 3:52–75
Bianchi V, Frascarelli M (2009) Is topic a root phenomenon? Ms, University of Siena, University of Roma 3
Bocci G (2009) On syntax and prosody in Italian. Doctoral dissertation, University of Siena
Bresnan J (1977) Variables in the theory of transformations. In: Culicover P et al (eds) Formal syntax. Academic Press, New York
Chomsky N (1981) Lectures on government and binding. Foris Publications, Dordrecht
Chomsky N (1995) The minimalist program. The MIT Press, Cambridge
Chomsky N (2000) Minimalist inquiries: the framework. In: Martin R et al (eds) Step by step—essays in minimalist syntax in honor of Howard Lasnik. The MIT Press, Cambridge
Chomsky N (2004) Beyond explanatory adequacy. In: Belletti A (ed) Structures and beyond: the cartography of syntactic structures, vol 3. Oxford University Press, New York
Cinque G (1999) Adverbs and inflectional heads. Oxford University Press, Oxford

Cinque G (ed) (2002) The structure of CP and DP: the cartography of syntactic structures, vol 1. Oxford University Press, Oxford
Cinque G, Rizzi L (2010) The cartography of syntactic structures. In: Heine B, Narrog H (eds) The Oxford handbook of linguistic analysis. Oxford University Press, Oxford, pp 51–65
Cruschina S (2008) Discourse-related features and the syntax of peripheral positions—a comparative study of Sicilian and other Romance languages. Doctoral Dissertation, University of Cambridge
Frascarelli M (2000) The syntax-phonology interface in focus and topic constructions in Italian. Kluwer Academic Publishers, Dordrecht
Frascarelli M, Hinterhölzl R (2007) Types of topics in German and Italian. In: Winkler S, Schwabe K (eds) On information structure, meaning and form. Benjamins, Amsterdam, pp 87–116
Haegeman L (1996) An introduction to government-binding theory. Blackwell, Oxford
Haegeman L (2011) Adverbial clauses, main clause phenomena, and the composition of the left periphery. Ms, University of Gent. To be published by Oxford University Press
Mulligan K (1987) Promisings and other social acts: their constituents and structures. In: Mulligan K (ed) Speech act and sachverhalt: reinach and the foundations of realist phenomenology. Nijhoff, Dordrecht, pp 29–90
Mulligan K (2006) Facts, formal objects and ontology, modes of existence. In: Bottani A, Davies R (eds) Papers in ontology and philosophical logic. Ontos Verlag, Frankfurt, pp 31–46
Rizzi L (1991) Residual verb second and the Wh criterion. Geneva working papers on formal and computational linguistics (republished in Rizzi 2000)
Rizzi L (1997) The fine structure of the left periphery. In: Haegeman L (ed) Elements of grammar. Kluwer, Dordrecht
Rizzi L (2000) Comparative syntax and language acquisition. Routledge, Oxford
Rizzi L (2004a) Locality and left periphery. In: Belletti A (ed) Structures and beyond: the cartography of syntactic structures, vol 3. Oxford University Press, New York
Rizzi L (ed) (2004b) The structure of CP and IP—the cartography of syntactic structures, vol 3. Oxford University Press, Oxford
Rizzi L (2006) On the form of chains: criterial positions and ECP effects. In: Cheng L, Corver N (eds) On Wh movement. The MIT Press, Cambridge
Rizzi L, Shlonsky U (2007) Strategies of Subject Extraction. In Gärtner H-M and Sauerland U (eds) *Interfaces + Recursion = Language? Chomsky's Minimalism and the View from Syntax-Semantics.* Mouton de Gruyter, Berlin, pp 115–116
Shlonsky, U (2010) The cartographic enterprise in syntax. *Language and Linguistics Com-pass*, 4(6):417–429

Kevin Mulligan's Bibliography

Publications

A Edited
B Articles
C Journalism, juvenilia, interviews
D Reviews
E Prefaces, introductions
F Short contributions to handbooks, encyclopedias
G Translations

A Edited

1987 ed. Speech Act and Sachverhalt: Reinach and the Foundations of Realist **Phenomenology**, Dordrecht: Nijhoff.

1990 ed. Mind, Meaning and Metaphysics: the Philosophy and Theory of Language of Anton Marty, Dordrecht: Kluwer.

1991 ed. **Language, Truth and Ontology**, Dordrecht: Kluwer.

1991 ed., issue of **Topoi,** *Continental Philosophy Analysed*, Vol. 10, No. 2.

1993 ed. (with J-P. Leyvraz) **Wittgenstein analysé**, Paris: Editions Jacqueline Chambon.

1993 ed. (with R. Roth) **Regards sur Bentham et l'utilitarisme**, *Recherches et Rencontres*, 4, Geneva: Droz.

1993 ed. (with Brian Garrett), **Themes from Wittgenstein**, *Working Papers in Philosophy*, 4, RSSS, Philosophy Program, Research School of Social Sciences, Australian National University, Canberra.

1999 ed. (with Patrizia Lombardo), special double number of **Critique**, *Penser les émotions*, 3/1999.

1999 ed. (with J.-P.Cometti), *La Critique de la raison en Europe centrale*, **Philosophiques**, 26/2. http://www.erudit.org/erudit/philoso/index.html.

2001 ed. (with J.-P. Cometti), La Philosophie autrichienne de Bolzano à Musil. Histoire et Actualité, Paris: Vrin.

2001 ed. (with B. Baertschi), **Les nationalismes**, "Ethique et philosophie morale", Paris: Presses Universitaires de France.

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8,

2010 Romanian tr. **Nationalismele**, Bucharest: Nemira.

2004 ed. (with H. Hochberg), **Relations and Predicates**, Philosophical Analysis, Frankfurt: Ontos Verlag, http://www.science-digital.com/Ontos/O-Mulligan-R.htm.

2007 ed. *The Philosophy of Kit Fine*, Special issue of **Dialectica**, 61, 1.

2009 ed. (with A. Westerhoff), **Robert Musil—Ironie, Satire, falsche Gefühle,** Paderborn: mentis Verlag.

2009 ed. (with Wlodek Rabinowicz), Special issue, *Value theory*, **Ethical Theory and Moral Practice**, 12, 4.

Forthcoming

2012 Mulligan, K., K. Kijania-Placek, K. & Placek, T., (eds.) Studies in the History and Philosophy of Polish Logic. Essays in Honour of Jan Woleński, Palgrave.

B Articles

1980 "Structure and Rules in Wittgenstein and Husserl", in **Language, Logic and Philosophy**. Proceedings of the IV. International Wittgenstein Symposium, 1979, 461–464.

1981 "Philosophy, Animality and Justice: Kleist, Kafka, Weininger and Wittgenstein", in ed. B. Smith **Structure and Gestalt: Philosophy and Literature in Austria-Hungary and her Successor States**, Amsterdam: Benjamins, 293–311.

1982 with B. Smith "Parts and Moments: Pieces of a Theory", in B. Smith ed. **Parts and Moments: Studies in Logic and Formal Ontology**, Munich: Philosophia, 15–109. http://wings.buffalo.edu/philosophy/faculty/smith/articles/pieces.htm.

1983 with B. Smith "Framework for Formal Ontology", in **Topoi**, 2, Special number on Lesniewski, 73–85, http://wings.buffalo.edu/philosophy/faculty/smith/articles/fffo.htm.

1983 "Spinoza on Necessary Existential Determination", in **Leibniz-Gesellschaft Leibniz-Werk und Wirkung**, Vorträge, IV. Internationaler Leibniz-Kongress, 532–539.

1983 "Constituency and Dependence in Language", in ed. Forskningsradsnamnden (Swedish Research Council). **An Inventory of Present Thinking about Parts and Wholes**, Vol. I, (Papers submitted for discussion), 75–89, 1983.

1983 "Colours and Complexity", in ed. Forskningsradsnamnden **An Inventory…**, Vol. II, Commentary, 43–50.

1984 with P. Simons, B. Smith "Truth-Makers", in **Philosophy and Phenomenological Research**, Vol. XIV, No. 3, 1984, 287–321, http://wings.buffalo.edu/academic/department/philosophy/faculty/smith/articles/truthmakers/tm.html.

German tr: 1987 "Wahrmacher", in Hgb. L. Bruno Puntel **Der Wahrheitsbegriff**, Darmstadt: Wissenschaftliche Buchgesellschaft, 210–255.

French tr.: 2011 "Vérifacteurs", *Études de philosophie*, no. 9–10, 2008–2011, translated by B. Langlet & J.-Fr. Rosecchi, pp. 104–138.

Reprint: 2007 J-M. Monnoyer (ed.) *Metaphysics and Truthmakers*, Frankfurt: ontos, 9–50.

Reprint: 2009 in: Lowe, E. J. and Rami, A. (eds.): *Truth and Truth-Making*, Chesham: Acumen, 59–86.

1985 "'Wie die Sachen sich zueinander verhalten' inside and outside the Tractatus", in **Teoria**, V/1985/2, *Wittgenstein and Contemporary Philosophy*, eds. B. McGuinness, A. Gargani.

1985 with B. Smith "Mach und Ehrenfels: Über Gestaltqualitäten und das Problem der Abhängigkeit", in Hgb. R. Fabian **Leben und Werk von Chr. v. Ehrenfels**, Bd. VII, ("Studien zur österreichischen Philosophie"), Amsterdam: Rodopi, 85–111.

Expanded English version: 1988 "Mach and Ehrenfels: On the Foundations of Gestalt Theory", in ed. B. Smith **Foundations of Gestalt Theory**, Munich: Philosophia, 124–157, http://ontology.buffalo.edu/smith/articles/mach/mach.html.

Rumanian tr.: 2005 "Mach si Ehrenfels. Fundamentele teoriei gestaltiste", in Constantin Stoenescu, Ion Tanasescu (eds.), *Filosofia Austriaca*, Bucharest: Pelican, 262–298.

1985 with R. Rug "Trieb und Theorie: Bemerkungen zu Ehrenfels und Freud", in Hgb.R. Fabian, **Leben und Werk von Chr. v. Ehrenfels**, Bd. VII, ("Studien zur österreichischen Philosophie"), Amsterdam: Rodopi, 214–246.

1986 "Exactness, Description and Variation—How Austrian Analytic Philosophy was Done", in Hgb. C. Nyiri **Von Bolzano zu Wittgenstein—Zur Tradition der österreichischen Philosophie**, Vienna: Holder-Pichler, 86–97.

1986 with B. Smith "A Relational Theory of the Act", in **Topoi**, V, Current Issues in Phenomenology, eds. A. Bonomi, D.Woodruff Smith, 115–130.

1986 with B. Smith "A Husserlian Theory of Indexicality", **Grazer Philosophische Studien**, (Chisholm Festschrift), 28, 133–163. http://wings.buffalo.edu/philosophy/faculty/smith/articles/indexica.htm.

1987 "Promisings and other Social Acts: their Constituents and Structures" in ed. K. Mulligan **Speech Act and Sachverhalt: Reinach and the Foundations of Realist Phenomenology**, Dordrecht: Nijhoff, 29–90.

Italian translation: 2000 "Promesse ed altri atti sociali: costituenti e struttura", eds. S. Besoli & L. Guidetti **Il Realismo fenomenologico**. Sulla filosofia dei circoli di Monaco e Gottinga, Macerata: Quodlibet, 309–384.

1988 "On the Notion of Structure: Bühler's Linguistic and Psychological Examples", in ed. A. Eschbach **Karl Bühler's Theory of Language**, Amsterdam: Benjamins, 203–226.

1988 "Seeing as and Assimilative Perception", **Brentano Studien**, I, 129–152.

1989 "Judgings: their Parts and Counterparts" in **Topoi Supplement**, 2, *La Scuola di Brentano*, 117–148.

1989 "The Expression of Exactness: Ernst Mach, the Brentanists and the Ideal of Clarity", in (ed.) Robert Pynsent, **Decadence and Innovation. Austro-Hungarian Life and Art at the Turn of the Century**, London: Weidenfeld & Nicolson, 33–42.

1990 "Genauigkeit und Geschwätz—Glossen zu einem paradigmatischen Gegensatz in der Philosophie" in Hgb. H. Bachmaier **Wien—Paradigmen der Moderne**, Amsterdam: Benjamins, 209–236.

French tr. 1999 "Exactitude et bavardage. Gloses pour une opposition paradigmatique dans la philosophie autrichienne", "La Critique de la raison en Europe centrale", **Philosophiques** (Canada), 26/2, 177–201, http://www.erudit.org/erudit/philoso/index.html.

Russian tr. Точность и болтовня *Глоссы к парадигматическим противопо ставлениям в австрийской философии*. http://www.ruthenia.ru/logos/number /2002_01/03.htm.

1990 "Husserl on States of Affairs in the **Logical Investigations**", **Epistemologia**, special number on *Logica e Ontologia*, XII, 207–234, (Proceedings of 1987 Genoa conference on Logic and Ontology).

Spanish tr. 1990 "Las situaciones objetivas en las **Investigaciones Logicas** de Edmundo Husserl", **Revista de Filosofia**, III/3, 23–49.

Revised Italian tr. 1997 "Lo stato di cose nelle *Ricerche Logiche* di Husserl", **Discipline filosofiche**, special number on **La nozione di "stato di cose"** (The notion of "state of affairs"), 2, 127–158.

1990 "Marty's Philosophical Grammar" in ed. K. Mulligan Mind, Meaning and Metaphysics: the Philosophy and Theory of Language of Anton Marty, 11–28.

1990 "Criteria and Indication", **Wittgenstein—Towards a Reevaluation**, Proceedings of the Kirchberg Wittgenstein Centenary Celebration, 1989, Vienna: Hölder-Pichler-Tempsky, 94–105.

1991 "How Not to Read: Derrida on Husserl", in *Continental Philosophy Analysed*, **Topoi,** Vol. 10, No. 2, 199–208.

Spanish tr. "Como no hay que leer: la crítica de Derrida a Husserl" (tr. by D. Caraoca), forthcoming.

1991 "Colours, Corners and Complexity: Meinong and Wittgenstein on some Internal Relations", in (eds.) B. C. van Fraassen, B. Skyrms & W. Spohn, **Existence and Explanation: Essays in Honor of Karel Lambert**, The University of Western Ontario Series in Philosophy of Science, Dordrecht: Kluwer, 77–101.

1993 "Internal Relations", **Working Papers in Philosophy**, 2, RSSS, Australian National University, Canberra, Proceedings of the 1992 Canberra metaphysics conference, (eds.) Brian Garrett & Peter Menzies, 1–22.

1993 "Proposizione, stato di cose e altri concetti formali nel pensiero di Wittgenstein e Husserl", **L'uomo, un segno**, Fascicolo speciale: **Wittgenstein contemporaneo**, a cura di A. Gargani, 41–65.

1993 "Description's Objects: Austrian Variations", eds. B. Garrett & K. Mulligan, **Themes from Wittgenstein**, Working Papers in Philosophy, 4, RSSS, Philosophy Program, Research School of Social Sciences, Australian National University, Canberra, 62–85.

1993 "Post-Continental Philosophy: Nosological Notes", in **Stanford French Review**, Special number: *Philosophy and the Analytic-Continental Divide*, 17.2–3, ed. Pascal Engel, 133–150 (appeared 1994).

Partial Italian tr. 1997 "Analitici e continentali; il pluralismo in filosofia": "1. Mulligan: la filosofia continentale dal punto di vista analitico", "2. La storia", "3. Fattori sociali", "4. Alle origini della filosofia continentale", F. D'Agostini, **Filosofia Analitica. Analizzare, tradurre, interpretare**, Turin: Paravia Scriptorium, 172–177.

French tr. 2000 "C'était quoi la philosophie dite 'continentale'?", K. O. Apel, J. Barnes et al. **Un siècle de philosophie 1900–2000**, Folio Essais, Paris: Gallimard, 332–366.

1995 "Perception", **Husserl. Cambridge Companions to Philosophy**, eds. B. Smith & D. Smith, Cambridge, 168–238.

1995 "Musils Analyse des Gefühls", in Hgb. B. Böschenstein & M.-L. Roth, **Hommage à Robert Musil**, (Proceedings of 1992 Geneva Musil conference), Berne: Lang, 87–110.

1995 "Le spectre de l'affect inverti et l'espace des émotions", *La Couleur des pensées*, eds. P. Paperman & R. Ogien, Paris: Editions de l'Ecole des Hautes Etudes en Sciences Sociales, **Raisons pratiques**, 6, 65–83.

English version: 1998 "The Spectre of Inverted Emotions and the Space of Emotions", **Acta Analytica**, 89–105.

1996 "Constancy, Content and Sense", **Penser l'Esprit. Des sciences de la cognition à une philosophie cognitive**, sous la dir. de V. Rialle et D. Fisette, Presses Universitaires de Grenoble, 141–150.

1997 "Konstanz und Kriterien: Brunswiks Beitrag", Hrsg. K. Fischer, F. Stadler, **Wahrnehmung und Gegenstandswelt. Zum Lebenswerk von Egon Brunswik**, Springer, Veröffentlichungen des Instituts Wiener Kreis, Bd. 4, (Proceedings of the 1994 Vienna Brunswik Colloquium), 137–150.

1997 "The Essence of Language: Wittgenstein's Builders and Bühler's Bricks", **Revue de Métaphysique et Morale**, 2, 193–216.

German tr. 1997 "Das Wesen der Sprache: Wittgensteins Maurer und Bühlers Bausteine", **Brentano Studien**, 7, 267–290.

French tr. 2004 "L'essence du langage, les maçons de Wittgenstein et les briques de Bühler", http://htl.linguist.jussieu.fr/num2/num2.htm.

Kodikas/Code. Ars Semeiotica, Special Issue, *Karl Bühler,* Vol. 28, 71–86.

1997 "Sur l'Histoire de l'approche analytique de l'histoire de la philosophie: de Bolzano et Brentano à Bennett et Barnes", (Ed.) J.-M. Vienne, **Philosophie analytique et Histoire de la philosophie** (Proceedings of 1991 Nantes conference), Paris: Vrin, 61–103.

1997 "How Perception Fixes Reference", in: Alex Burri (ed.), **Sprache und Denken/Language and Thought**, Berlin/New York: de Gruyter, 122–138.

1998 with R. Mulligan & A.-C. Juillerat, "La Mémoire affective: le cas de la douleur", **Cahiers de psychiatrie**, Genève, 157–161.

1998 "From Appropriate Emotions to Values", *Secondary Qualities Generalized*, ed. P. Menzies, **The Monist**, vol. 84, no. 1, January, 161–188. http://www.unige.ch/lettres/philo/enseignants/km/doc/FromAppropriateEmotions.pdf.

German tr. 2009 "Von angemessenen Gefühlen zu Werten", ed. Sabine Döring, **Philosophie der Gefühle**, Frankfurt am Main: Suhrkamp Verlag, 462–495.

1998 "Relations—through thick and thin", **Erkenntnis**, *Analytical Ontology*, 325–353. French tr. forthcoming, Vrin: Paris.

1999 "Justification, Rule-Breaking and the Mind", **Proceedings of the Aristotelian Society**, London, Vol. XCIX, 123–139.

1999 "Perception, Predicates and Particulars", ed. Denis Fisette, **Consciousness and Intentionality: Models and Modalities of Attribution**, The Western Ontario Series in Philosophy of Science, Kluwer, 163–194.

Abbreviated Spanish version: 1996 "Percepción, Particulares y Predicados", **Revista de Filosofia**, III epoca, 9, 105–120.

1999 "La varietà e l'unità dell'immaginazione", **Rivista di Estetica**, *Percezione*, 53–67.

2000 "Métaphysique et Ontologie", (dir.) P. Engel, **Précis de Philosophie analytique**, Collection Thémis, Paris: Presses Universitaires de France, 5–33.

Italian tr., 2002, "Metafisica e ontologia", **Aut Aut**, 310–311, 116–143.

2000 "La Philosophie autrichienne—quelques variations constantes", Ouelbani, M (dir.) La Philosophie autrichienne—spécificités et influences, Université de Tunis 1, 9–26.

Expanded version: 2001 "De la philosophie autrichienne et de sa place", J.-P. Cometti & K. Mulligan eds., **La Philosophie autrichienne de Bolzano à Musil**. **Histoire et** Actualité, Paris: Vrin, 8–25.

2001 "Actes i objectes. Una anàlisi de la fenomenologia realista", Anuari de la societat catalana de filosofia, XIII, Institut d'estudis catalans, 241–262.

Steven Hales, Wadsworth: Belmont, California, 83–92.

2002 "Getting Geist—Certainty, Rules and Us", **Cinquantenaire Ludwig Wittgenstein**, Proceedings of the 2001 Tunis Wittgenstein conference, ed. M. Ouelbani, University of Tunis, 35–62.

2003 "Seeing, Certainty and Apprehension", Proceedings of the 2000 Melbu Conference on Non-Conceptual Content, Hallvard Fossheim, Tarjei Mandt Larsen, and John Rickard Sageng (eds.), **Non-Conceptual Aspects of Experience**. Oslo: Unipub forlag, 2003. ISBN: 82-7477-118-4, 27–44.

2003 with Pascal Engel, "Normes éthiques et normes cognitives", **Cités**, N°15, 2003, PUF, Paris, pp. 171–186.

2003 "Searle, Derrida and the Ends of Phenomenology", **John Searle**, ed. B. Smith, Contemporary philosophy in Focus, Cambridge University Press, 261–286.

2003 "Dispositions, their Bases and Correlates—Meinong's Analysis", **Philosophy and Logic. In Search of the Polish Tradition**, ed. Katarzyna Kijania-Placek, Synthese Series, FS for Jan Wolenski, Kluwer, 193–211.

2003 "Stati di cose, verità e fattori di verità", special number of *Sistemi intelligenti*, on Ontology, ed. R. Casati, XV, 3, (translated by Alessandro Dell'Anna), 539–556.

2003 "La filosofia analitica: che cosa è stata e che cosa ha da essere", *Iride*, 40, 631–634.

2003 "Forms of Life or Ways of Life?", *Rivista di estetica*, 24, 3/2003, XLIII, "Bozetti. In Memoria di Paolo Bozzi", eds. C. Barbero, R. Casati, M. Ferraris, A. Varzi, 103–105.

2004 "Brentano on the Mind", **Cambridge Companion to Brentano**, ed. D. Jacquette, Cambridge University Press, 66–97.

2004 "Essence and Modality. The Quintessence of Husserl's Theory", in M. Siebel & M. Textor (eds.) **Semantik und Ontologie. Beiträge zur philosophischen Forschung**, Frankfurt: Ontos Verlag, 387–418.

2004 "Husserl on the 'logics' of valuing, values and norms", *Fenomenologia della Ragion Pratica. L'Etica di Edmund Husserl*, (eds.) B. Centi & G. Gigliotti, 177–225, Naples: Bibliopolis.

French tr.: 2006 "Husserl sur les 'Logiques' de la valorisation, des valeurs et des norms", *Philosophia Scientiae*, 10, I, 71–107.

2006 "Soil, Sediment and Certainty", **The Austrian Contribution to Analaytic Philosophy,** ed. Mark Textor, London: Routledge (London Studies in the History of Philosophy), 89–129.

2006 "Wahrheit und das Wahrmacher-Prinzip im Jahr 1921", **Untersuchungen zur Ontologie**, (eds.) G. Imaguire & C. Schneider, Munich: Philosophia, Festschrift for Hans Burkhardt, 55–78.

2006 "Facts, Formal Objects and Ontology", **Modes of Existence. Papers in Ontology and Philosophical Logic**, eds. Andrea Bottani & Richard Davies, Frankfurt: ontos verlag, 31–46. http://www.unige.ch/lettres/philo/enseignants/km/doc/FactsBergamo4.pdf.

2006 with Peter Simons & Barry Smith, "What's Wrong with Contemporary Philosophy?", special number of **Topoi**, *Philosophy: What is to be done?*, 25, 63–67. http://www.springerlink.com/content/e6hl522358431760/.

2006 "Geist (and Gemüt) vs Life—Max Scheler and Robert Musil", **Le Ragioni del Conoscere e dell'Agire. Scritti in onore di Rosaria Egidi,** ed. R. Calcaterra, Milan: Franco Angeli, 366–378.

2006 "Ascent, Propositions and other Formal Objects", **Grazer Philosophische Studien**, 72, 29–48. http://www.unige.ch/lettres/philo/enseignants/km/doc/StatesAffPropPadova4.pdf.

2007 "Intentionality, Knowledge and Formal Objects", **Hommage à Wlodek**, electronic Festschrift for Wlodek Rabinowiz, eds. Toni Rønnow-Rasmussen et al. http://www.fil.lu.se/HommageaWlodek/site/papper/MulliganKevin.pdf.

Expanded print version, 2007 "Intentionality, Knowledge and Formal Objects", **Disputatio**, Vol. II, No. 23, November 2007, Special Number, 205–228, http://disputatio.com/index.php.

2007 "Two Dogmas of Truthmaking", J-M. Monnoyer (ed.) **Metaphysics and Truthmakers**, Frankfurt: ontos, 51–66. http://www.unige.ch/lettres/philo/enseignants/km/doc/TMTwoDogmas.pdf.

2008 "Ironie, valeurs cognitives et bêtise" (French tr. of "Irony, Cognitive Values and Foolishness"), *Philosophiques*, Vol. 35, No. 1, *Les valeurs de l'ironie*, ed. Pascal Engel, 89–107, http://www.unige.ch/lettres/philo/enseignants/km/doc/ironie_betise.pdf.

2008 "Scheler: Die Anatomie des Herzens oder was man alles fühlen kann", **Klassische Emotionstheorien von Platon bis Wittgenstein** (Classic Theories of Emotions from Plato to Wittgenstein), eds. H. Landweer & U. Renz, Berlin: de Gruyter, 589–612.

2008 "Propriétés, Processus et Priorités", *Compléments de substance. Etudes sur les propriétés accidentelles offertes à Alain de Libera*, eds. Ch. Erismann & A. Schniewind, Paris: Vrin, 231–247.

2009 "Truth and the truth-maker principle in 1921", in: Lowe, E. J. and Rami, A. (eds.): *Truth and Truth-Making,* Chesham: Acumen, 39–58. http://www.acumenpublishing.co.uk/display.asp?K=e2008020112552008&sf1=editor&st1=E%20J%20Lowe%20and%20A%20Rami&sort=sort_title&m=1&dc=1.

2009 "Selbstliebe, Sympathie, Egoismus", **Robert Musil—Ironie, Satire und falsche Gefühle**, eds, K. Mulligan & A. Westerhoff, 55–73. http://www.unige.ch/lettres/philo/enseignants/km/doc/Selbstliebe2.pdf.

2009 "Was sind und was sollen die unechten Gefühlen?", Hgb. Ursula Amrein, **Das Authentische. Referenzen und Repräsentationen**, Zürich: Chronos Verlag (www.chronos- verlag.ch), 225–242.

2009 "On Being Struck by Value—Exclamations, Motivations and Vocations", **Leben mit Gefühlen. Emotionen, Werte und ihre Kritik**, ed. Barbara Merkel, Paderborn: mentis-Verlag, 141–161.

2009 "Torheit, Vernünftigkeit und der Wert des Wissens", **Wissen und Werte**, ed. G. Schönrich, Paderborn: mentis Verlag, 27–44.

2009 "Tractarian Beginnings and Endings. Worlds, Values, Facts and Subjects", in: G. Primiero, S. Rahman (eds), **Acts of Knowledge: History, Philosophy and Logic**. **Essays Dedicated to Göran Sundholm**, College Publications, Tribute series, 151–168.

2009 "Values", **The Routledge Companion to Metaphysics**, eds. R. Poidevin, P. Simons, A. McGonigal & R. Cameron, London: Routledge, 401–411.

2010 "Emotions and Values", **Oxford Companion to the Philosophy of Emotions**, ed. P. Goldie, Oxford University Press, 475–500.

2010 "The Truth Connective vs the Truth Predicate. On Taking Connectives Seriously", part of a symposium on Wolfgang Künne's *Conceptions of Truth* (with a reply by Künne), *Dialectica*, 64, 4, 565–584, http://onlinelibrary.wiley.com/doi/10.1111/j.1746-8361.2010.01247.x/pdf.

2010 "Husserls Herz", *Husserl und die Philosophie des Geistes*, eds. Manfred Frank & Niels Weidtmann, Berlin: Suhrkamp, 209–238.

2011 "On Meaning Something and Meanings", **Themes from Early Analytic Philosophy. Essays in Honour of Wolfgang Künne**, special number *Grazer Philosophische Studien*, ed. B. Schnieder et al., 82, 255–284.

2011 Hastings, J., Ceusters, W., Smith, B., Mulligan, K. "Dispositions and Processes in the Emotion Ontology" Proceedings of the International Conference on Biomedical Ontology (ICBO), 26–30 July 2011. http://icbo.buffalo.edu/ICBO-2011_Proceedings.pdf.

2011 **Wittgenstein et ses prédecesseurs austro-allemands**, Conférences Hugues Leblanc, Montreal, I "De l'esprit et de l'âme", II "Eprouver vs Vouloir Dire, Vouloir, Se Souvenir", III "Significations primaires et secondaires", *Philosophiques*, 38, 2, pp. 5–69. http://www.erudit.org/revue/philoso/2011/v38/n1/.

Forthcoming or submitted

2012 with Scherer, K. R. "Toward a Working Definition of Emotion", *Emotion Review*, with peer commentary.

2012 "William James meets his 'German' Critics".

2012 "Acceptance, Acknowledgment, Affirmation, Agreement, Assertion, Belief, Certainty, Conviction, Denial, Judgment, Refusal & Rejection", for a volume ed. by M. Textor.

2012 "Czesław Miłosz et les valeurs cognitives", *Czesław Miłosz, l'Europe et la Russie*

2012 "Formal Concepts—from Bolzano & Husserl to Wittgenstein & Tarski", eds. Mulligan, K., K. Kijania-Placek, K. & Placek, T., *Studies in the History and Philosophy of Polish Logic. Essays in Honour of Jan Woleński*, Palgrave.

2012 "Husserls Phantasien", to appear in a volume edited by W. Künne, V. Klostermann.

2012 "Husserls Iche", to appear in a volume edited by W. Künne, V. Klostermann.

2012 "The Meanings of *Bedeutung*", *Between Mind and Language. Anton Marty and Karl Bühler*, L. Cesalli & J. Friedrich, eds., Basel: Schwabe.

2012/3 Gerrans, Ph. & Mulligan, K., "Intentional Imagination and Delusion".

C Journalism, juvenilia, interviews

1978 "Inscriptions and Speaking's Place", **Oxford Literary Review**, vol. 3, no. 2, 62–67.

1982 Contributor with P.Simons to: B.Smith "Annotated Bibliography of Writings on Part-Whole Relations since Brentano (with special reference to Psychology and Linguistics", in ed. B. Smith **Parts and Moments**, 481–552.

1984 "Anton Marty: ein Schweizer Philosoph in Prag—Rückblick auf den Sprachforscher", in **Neue Zürcher Zeitung**, 29/XI/84, 37–38.

1986 "Interview avec Kevin Mulligan", Bulletin du groupe genevois de la société romande de philosophie, 4, 3–8.

1987 with P. Simons and B. Smith, "Drei Briten in Kakanien", Interview in **Information Philosophie**, 3, 22–33, http://www.unige.ch/lettres/philo/enseignants/km/doc/Kakanien.pdf.

1991 "De la nécessité vitale de la transparence", **Campus**, (Geneva), 9, 8–9.

1998 "Valeurs et Normes Cognitives", *Les nouvelles morales—éthique et philosophie*, ed. Monique Canto-Sperber, **Magazine Littéraire**, 78–79, http://peccatte.rever.fr/SBPresse/MagazineLitteraireKM011998.html.

1998 "The symptoms of Gödel-mania. Parisian abuses of science and postmodernist discourse", **Times Literary Supplement**, review of A. Sokal & J. Bricmont **Les Impostures Intellectuelles,** 1.5.98, 13–14; http://naturalscience.com/ns/books/book04.html.

1998 "The great divide", **Times Literary Supplement,** (Title page: "The battle of the two schools") review of books by Engel, d'Agostini and others, 26.6.98, 6–8.

1998 "Continental and analytic philosophy", Letter to the Editor, **Times Literary Supplement**, 24.07.98, p. 17.

2001 "Kevin Mulligan. Subversão e filosofia", Interview (Desidério Murcho), Livros, 18 April, 54–55, http://critica.no.sapo.pt/entr_kmulligan.html.

2001 "Wittgenstein and Austrian Philosophy", Television Interview, RAI Educational—Multimedia Encyclopaedia.

D Reviews

1973 with Karl-Peter Markl & Ali Rattansi, "Full Marx" [Review of books by P. Walton & S. Hall and by J. O'Malley], **Radical Philosophy**, 4, Spring, 41–43.

1984 with B. Smith "Traditional vs. Analytic Philosophy". Critical Notice of E. Tugendhat "Traditional and Analytic Philosophy" in "**Grazer Philosophische Studien**" Vol. 21, 193–2O2.

1985 with B. Smith "Franz Brentano on the Ontology of Mind". Critical Notice of F. Brentano "Deskriptive Psychologie" in **Philosophy and Phenomenological Research**, Vol. XLV, No. 4, 627–644, http://ontology.buffalo.edu/smith//articles/brentano/ontology_of_mind.pdf.

1986 with B. Smith Critical Notice of E. Husserl "Logische Untersuchungen", Husserliana XIX/1–2, Hgb. U. Panzer, in **Grazer Philosophische Studien**, 27, 199–207.

1990 "Amore della perspicuità", Indice, review of Wittgenstein's Osservazioni sulla filosofia della psicologia (= Italian translation of Bemerkungen über die Philosophie der Psychologie), p. 4, June.

1995 "Psychologism and its History Revalued" (Review of Martin Kusch, Psychologism: A Case Study in the Sociology of Knowledge), **Metascience**, 8, 17–26, with reply by Kusch.

1999 [Review of **Austrian Philosophy Past and Present. Essays in Honour of Rudolf Haller**, eds K. Lehrer & J. C. Marek, Boston Studies in the Philosophy of Science, 190, Kluwer, 1997], **Institute Vienna Circle Yearbook**, Greenberger-Reiter-Zeilinger (eds.): *Epistemological and Experimental Problems in Quantum Physics*, 350–353.

2000 [Review of Busino, G 1998 **Sociologie des sciences et des techniques**, Paris: PUF, Que sais-je?], **European Societies**, 2 (1), 101–103.

2011 "Quibbles & Grumbles from Mitteleuropa", part of a symposium on Hans-Johann Glock's *What is Analytic Philosophy?*, with replies by Glock, *Teorema*, XXX/1, 103–113.

E Prefaces, introductions

1987 "Preface" to ed. K. Mulligan Speech Act and Sachverhalt: Reinach and the Foundations of Realist Phenomenology, Dordrecht: Nijhoff, vii.

1990 "Preface" to ed. K. Mulligan Mind, Meaning and Metaphysics: the Philosophy and Theory of Language of Anton Marty, xi.

1991 "Introduction: On the History of Continental Philosophy", in **Continental Philosophy Analysed**, 115–120.

Partial Italian tr. 1992 "Sulla Storia e l'Analisi della Filosofia Continentale", **Iride**, 8, 183–190.

1990 with Karl Schuhmann "Two Letters from Marty to Husserl", in ed. K. Mulligan **Mind, Meaning and Metaphysics: the Philosophy and Theory of Language of Anton Marty**, 225–236.

1991 "Preface" to (ed. K. Mulligan) **Language, Truth and Ontology**, ix–x.

1999 with Patrizia Lombardo, "Avant-propos", **Critique**, *Penser les émotions*, 3, 481–486.

1999 with J.-P.Cometti (dirs.) "Introduction", "La Critique de la raison en Europe centrale", **Philosophiques** (Canada), 26/2. http://www.erudit.org/erudit/philoso/index.html.

2000 with J.-P. Cometti, "Introduction" in K. Mulligan & J.-P. Cometti, eds. **La Philosophie autrichienne**, Proceedings of the 1997 Cerisy colloquium.

2001 with B. Baertschi, "Avant-propos", Baertschi & Mulligan, eds., **Les nationalismes**, "Ethique et philosophie morale", Paris: Presses Universitaires de France, 1–2.

2001 with B. Baertschi, "Introduction", Baertschi & Mulligan, eds., **Les nationalismes**, "Ethique et philosophie morale", Paris: Presses Universitaires de France, 3–8.

2010 Romanian tr. "Introducere", **Nationalismele**, Bucharest: Nemira, 8–14.

2003 "Wittgenstein analysé", Preface to **Wittgenstein analysé**, 5–12.

2004 with R. Roth "Présentation", **Regards sur Bentham et l'utilitarisme** (Actes du colloque organisé à Genève les 23 et 24 novembre 1990), Geneva: Droz, 7–9.

2005 with B. Garrett, "Preface", eds. B. Garrett & K. Mulligan, **Themes from Wittgenstein**, Working Papers in Philosophy, 4, RSSS, Philosophy Program, Research School of Social Sciences, Australian National University, Canberra, i–ii.

2009 with Armin Westerhoff: "Statt einer Einleitung: Drei Stichworte und zwei Kontexte zu Robert Musil", **Robert Musil—Ironie, Satire und falsche Gefühle**, eds, K. Mulligan & A. Westerhoff, 7–11.

F Contributions to handbooks, encyclopedias

1992 "Gardies, Jean-Louis…Esquisse d'une grammaire pure", **Encyclopédie Philosophique Universelle**, III, **Les Oeuvres Philosophiques**, Tome 2, Paris: Presses Universitaires de France, p. 3252.

1995 "Inherence" in **A Companion to Metaphysics**, eds. E. Sosa & J. Kim, Oxford: Blackwell, 242–243.

1995 "Internal Relations" in **A Companion to Metaphysics**, eds. E. Sosa & J. Kim, Oxford: Blackwell, 245–246.

1995 "Mach" in **A Companion to Metaphysics**, eds. E. Sosa & J. Kim, Oxford: Blackwell, 87–88.

1995 "Relation", in **A Companion to Metaphysics**, eds. E. Sosa & J. Kim, Oxford: Blackwell, 445–446.

1998 "Predication", **Routledge Encyclopaedia of Philosophy**, (ed.) E. Craig, Routledge: London, Vol 7, 665–667.

Extract: 2000 "Predication", **Concise Routledge Enyclopedia of Philosophy**, ed. E. Craig, Routledge, 708.

2001 "Logical Positivism and Logical Empiricism", N. J. Smelser and Paul B. Baltes (editors), **International Encyclopedia of the Social & Behavioral Sciences**, Pergamon, Oxford, 9036–9038.

2001 "Phenomenology: Philosophical Aspects", N. J. Smelser and Paul B. Baltes (editors) **International Encyclopedia of the Social & Behavioral Sciences**, Pergamon, Oxford, 11363–11369.

Catalan tr.: 2002 "Actes socials i objectes socials", **Comprendre. Revista catalana de filosofia**, Any IV, 2002/2, 193–204.

2007 with Fabrice Correia "Facts", **Stanford Encyclopedia of Philosophy**, http://plato.stanford.edu/entries/facts/.

2009 "Moral Emotions", in eds. Sander, D. & Scherer, K., **The Oxford Companion to Emotions and the Affective Sciences**, Oxford University Press, 262–264. http://www.unige.ch/lettres/philo/enseignants/km/doc/EmotionMoral.pdf.

2009 "Gestalt (and feeling)", in eds. Sander, D. & Scherer, K., **The Oxford Companion to Emotions and the Affective Sciences**, Oxford University Press, 195–196.

G Translations

1978 with Th. Loff, translation of H.-J. Heringer Praktische Semantik: Practical Semantics: a Study in the Rules of Speech and Action, Mouton.

1982 translation of R. Musil **Beitrag zur Beurteilung der Lehren Machs: On Mach's Theories**, Introduction by G.H. von Wright Munich: Philosophia.

1985 translation of J.-L. Gardies **Esquisse d'une Grammaire Pure**, (with additional chapters on adverbs, names of states of affairs and aspect): **Rational Grammar**, Munich: Philosophia.

Index

A. Reboul (ed.), *Mind, Values, and Metaphysics,* DOI 10.1007/978-3-319-05146-8,