Hans Bernhard Schmid

Contributions to Phenomenology 58

Plural Action

Essays in Philosophy and Social Science

Springer

PLURAL ACTION

CONTRIBUTIONS TO PHENOMENOLOGY

IN COOPERATION WITH
THE CENTER FOR ADVANCED RESEARCH IN PHENOMENOLOGY

Volume 58

Scope

The purpose of the series is to serve as a vehicle for the pursuit of phenomenological research across a broad spectrum, including cross-over developments with other fields of inquiry such as the social sciences and cognitive science. Since its establishment in 1987, *Contributions to Phenomenology* has published nearly 60 titles on diverse themes of phenomenological philosophy. In addition to welcoming monographs and collections of papers in established areas of scholarship, the series encourages original work in phenomenology. The breadth and depth of the Series reflects the rich and varied significance of phenomenological thinking for seminal questions of human inquiry as well as the increasingly international reach of phenomenological research.

PLURAL ACTION

Essays in Philosophy and Social Science

by

HANS BERNHARD SCHMID

Springer

Hans Bernhard Schmid
University of Basel
Department of Philosophy
Nadelberg 6-8
CH-4051 Basel
Switzerland
Hans-Bernhard.Schmid@unibas.ch

ISSN 0923-9545
ISBN 978-90-481-2436-7 e-ISBN 978-90-481-2437-4
DOI 10.1007/978-90-481-2437-4
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009926505

Acknowledgements

Working on a topic such as the one of this book cannot possibly be a solitary venture. There are many people and institutions from which I have received support while working on this project. Let me just mention the following. First and foremost, I wish to thank Raimo Tuomela. In his 1984 book, and especially in his paper on *We-Intention* (1988; co-authored with Kaarlo Miller), he set the stage for what was to come in the 2 decades that have followed. Moreover, he has been a *spiritus rector* of the analysis of collective intentionality not only with his own many subsequent contributions but also by bringing those working in the field together.

Anita Konzelmann Ziv and Juliette Gloor have been working together with me for more than 2 years now. I am grateful to them for their comments. Also, I wish to thank Rebekka Gersbach and Samuel Häfner for their help in preparing the manuscript for print, and to Jeff Kochan and Barbara Hauenstein for checking the proofs. Two anonymous referees for Springer made sharp and helpful comments on an earlier draft; many thanks to them. Last but not least I am grateful to the Swiss National Science Foundation for providing excellent working conditions.

Earlier versions or parts of the chapters in this book have appeared in the following journals and collected volumes (by chapter number):

1. *Philosophy of Social Science* 38/1 (2008), pp. 25–54.
2. *Philosophical Explorations* 6/3 (2003), pp. 201–218.
3. Nikos Psarros/K. Schulte-Ostermann (eds.), *Facets of Sociality. Holistic Epistemology and Action Theory*, Heusenstamm: Ontos 2007, pp. 299–317.
4. Hans Bernhard Schmid/K. Schulte-Ostermann/N. Psarros (eds.), *Concepts of Sharedness.* Frankfurt: Ontos 2008, pp. 59–86.
5. *Analyse & Kritik* 27/1 (2005), Symposium on Ernst Fehr: The Nature of Human Altruism, pp. 172–188.
6. Mark D. White/Barbara Montero (eds.), *Economics and the Mind.* London: Routledge 2006, pp. 159–179.
7. *Economics & Philosophy* 21/1 (2005), pp. 51–63.
8. *Journal of the British Society for Phenomenology* 35/2 (2004), pp. 132–156.
9. *Dialektik – Zeitschrift für Kulturphilosophie* 16 (2005), Heft 1, pp. 157–170 (in German).

10. *Distinction. Scandinavian Journal for Social Theory* 9 (2004), pp. 103–118.
11. Brigitte Hilmer/Georg Lohmann/Tilo Wesche (eds.), *Anfang und Grenze des Sinns.* Weilerwist: Velbrück 2006, pp. 236–250 (in German).

I wish to thank the copyright holders for their kind permission to reuse this material.

Contents

List of Figures

Introduction

Collective intentionality is a label for shared attitudes of any kind – cognitive (shared belief), conative (shared intention), or affective (shared emotion).[1] Among these basic types of collective intentional states, joint intentions play an especially important role, and have been the focus of the analysis of collective intentionality ever since the label first came into use.[2] Collective intention is essential for such basic social phenomena as coordination, cooperation and communication.

Opening up a new perspective on the basic structure of the social world, the analysis of collective intentionality is one of the most conspicuous recent developments in philosophical research. Over the last 2 decades, this field of study has attracted considerable attention from a wide range of philosophical sub-disciplines, such as social ontology, the theory of practical reason, the philosophy of social science, and the ethics of collective responsibility, as well as from neighboring fields such as social theory, cognitive science, economic theory, linguistics, and developmental psychology, where some of the conceptual tools developed in collective intentionality analysis are already in use.

This volume contributes to this rapidly evolving research program by pursuing three basic aims. The first is conceptual. In the following, some of the main conceptual problems in the analysis of collective intention are introduced, and some possible solutions are suggested (Part I). Second, a number of examples are given for the use of collective intentionality analysis in the theory and philosophy of the social sciences (Part II). Third, it is shown how this line of research opens up new perspectives on classical topics in the history of social philosophy and social science, and how, conversely, an inquiry into the history of ideas can lead to further refinement of our conceptual tools in the analysis of collective intentionality (Part III).

The following is a summary of the main ideas and arguments developed in this volume. The first chapter sets the stage by introducing the concept of plural action.

---

[1] Other types of shared attitudes analyzed in the current debate include joint attention and joint acceptance.

[2] The term was coined by John Searle in his seminal paper on Collective Intentions and Actions (1990). Searle's concept of collective intentionality is very similar to Sellars' concept of *we-intention* (Sellars 1974, 1980, 1992). Sellars concept has its roots in George Robin Collingwood's use of the term in his *New Leviathan* (1947 [1942]).

Plural actions are defined as a kind of *social action.* Social actions are actions that require the activity of more than one individual (either as a matter of conceptual necessity, or contingently). Social action can either be *singular* or *plural.* In the case of social actions of the *singular* kind, the activity of the participating individuals is directed towards *different* goals, whereas in the case of plural actions, the participants *share* a goal in that they aim at achieving their goal *together*.

Conceptually, actions require agents, and this is where the trouble with plural actions starts. For somebody to be an agent, there has to be a description under which she *intended* to do what she did. It seems plausible, however, that in a core sense of the word, intention is *action self-referential*, i.e. agents can only intend what they take to be *their own* actions. As plural actions require the activity of *many*, the question of the plural agent arises: who could be so self-confident as to take *herself* to be doing what requires the activity of *many* to do? Ordinary language offers three candidates for that role. Plural actions are either ascribed to *collective agents* (such as in the case of Parliament being ascribed the action of passing a law), or to *influential individuals* (such as in the case of Caesar's being credited with defeating the *Helvetii* at *Bibracte*), or to the *joint activity of teams* (such as in the case of a couple of friends going for a walk together). Each of these candidates, however, meets with considerable skepticism from the side of the received theory of action. The first candidate – the collective agent – ruffles the feathers of methodological individualists who claim that the only suitable candidates for the role of agents are individuals. I argue in the first chapter that the basic worry behind the individualistic view is that a robust conception of collective agency somehow compromises or even displaces the agency of the participating individuals. Thus methodological individualism is basically a commitment to what I propose to call *individual intentional autonomy,* which is the assumption that the individual participants in plural actions *are agents,* i.e. that it is possible to interpret their behavior as being *their own action*. As recent conceptions of collective agency show, however, the worry that individual intentional autonomy is incompatible with a robust conception of collective agency is unfounded. Thus there is no reason not to accept a robust notion of collective agency.

In recent literature, the constitution of collective agents has received a considerable amount of attention, especially in Philip Pettit's work, and in Margaret Gilbert's *Plural Subject Theory.* There are, however, reasons for doubt concerning the scope of collective subject explanations of plural actions. First, only *very few* cases of plural action are fit candidates for collective agent explanations. Second, and more importantly, the constitution of a plural agent is itself a plural action. Thus the theory of plural action cannot stop there. Plural agents as collective agents presuppose plural agents of another kind. Therefore, the remaining two candidates proposed for the role of a plural agent by ordinary language need to be examined. The first of these candidates, however, seems even more problematic than collective agents. It is not without reason that the ordinary-language tendency to ascribe plural actions to individual leaders is met with reserve from the side of current humanities and social science. There is an air of conceptual confusion as well as of political incorrectness about such ascriptions. If plural actions are ascribed to only *one* of

the many participating individuals, it seems that the agency of that individual is unfairly extended at the cost of the agency of the other participants. It is as if the leader were taken to be the true agent behind the other participants' behavior, which seems incompatible with the view that, in plural actions, *all* participating individuals are agents in their own right. The more fundamental question behind this problem is this: what role can one individual's pro-attitudes play in the interpretation of another individual's behavior? There is a widespread tendency to endorse what I propose to call *motivational individualism*, or *individual motivational autarky.* This is the view that all intentional interpretations of an individual's behavior have to bottom out in that individual's *own* volitions, or pro-attitudes. I argue, first, that the main reason why the assumption of motivational autarky is endorsed is that it is not sufficiently distinguished from intentional autonomy. After a closer examination of the relation between these claims, I argue, secondly, that it is possible to *reject* the autarky claim while holding on to the autonomy assumption, and finally, that there are serious reasons for doubt concerning the autarky assumption.

Thus there is no reason not to allow concepts such as influence, power, and authority to play a greater role in action theory, and to allow for cases in which plural actions are ascribed to the agency of powerful individuals. It is clear, however, that just like in the case of the first type, this second type of plural agency applies only to a minor subset of the total class of plural actions, and that many of these cases presuppose plural agency of yet another type.

The last and most fundamental type of plural agent – I propose to call it the *teamwork model* – finally brings us to the question of collective intentionality. In teamwork explanations, plural actions are neither credited to the agency of a collective *qua* entity that is somehow *different* from the participating individuals, as in the collective agent model, nor to the agency of some leading individual participant, but rather to the *joint agency of many individuals*. I argue that the teamwork model is more fundamental than the collective agent model and the leadership model in that the latter two presuppose the former. It is by virtue of teamwork that plural agents and individual power positions come about.

The trouble with this model is how to reconcile the supposed *unity of action* with the *plurality of agents*. It follows from the assumption of individual intentional autonomy that agents perform *their own* actions. This, however, seems to preclude the possibility of many agents performing *one and the same* (token) action: if A's action and B's action are the same, A performs B's action. At the same time, however, we routinely ascribe actions to teams, without assuming a collective agent over and above the heads of the individual participants, or asymmetric power relations of the kind at work in individual leadership. So how, then, can the unity of action be reconciled with the plurality of agents without scratching the principle of individual intentional autonomy?

This is where the concept of collective intentionality comes into the picture. It is assumed in the current debate that many agents can indeed perform one (token) action precisely insofar as they *share the respective intention.* This, however, raises the question of how something like an intention can be shared. What can the word "sharing", as applied to mental states, possibly mean? In ordinary language, we

share such things as cars and cakes, and in these cases, "sharing" seems to imply one (token) cake with many pieces, or one (token) taxi with many passengers. In contrast to this straightforward sense of the word, however, most authors in the current debate on collective intentionality use a *metaphorical sense* of "sharing", which conforms to what I propose to call *intentional individualism*. Here, it is assumed that when people "share" an intention, each of them has his or her own intention, but that these individual intentions are of the same *type* (and perhaps accompanied by some reciprocal cognitive attitudes). So (token) intentions are not *literally* shared after all. This is what I call the *distributive* reading of collective intentionality, and one of the main aims of this volume is to cast doubt on this reading and defend the straightforward sense of "sharing" a (token) intention. In the first chapter, I present and discuss two arguments for the straightforward view of sharing. First, I argue that distributive readings of collective intentionality are either too broad (two people having an intention of the same *type* do not necessarily intend to act *together*), or viciously circular (this latter point is further pursued in Chapter 2). Second, I argue that the main reason why philosophers of collective intentionality endorse a distributive reading is that they think that intentional autonomy entails intentional individualism. I argue that this assumption is mistaken.

In the second chapter I examine the role of a rather somber figure that haunted the early stages of the debate on collective intentionality: the specter of the group mind. The group mind plays a crucial and fateful role in how the leading figures in the debate originally conceived of collective intentionality. Fear of the group mind is one important reason why those philosophers resorted to (intentional) individualism. My aim in this chapter is not to defend the group mind (the history of the most frightening version of the group mind, the *Volksgeist*, is examined in Chapter 10). Rather, I argue that fear of the group mind has done more damage than good to the philosophy of collective intentionality, and that it is time to take a more relaxed stance on the matter.

I examine how two of the most important philosophers of collective intentionality, i.e. John R. Searle and Michael E. Bratman, try to exorcise the group mind by resorting to two different versions of individualism, a venture within which they come to contrary views concerning the basic intentional structure of joint action. I argue in this chapter that both versions of individualism offer inadequate accounts of the structure of joint action and should thus be rejected, and also that it is as unnecessary as it is detrimental to our understanding of what it means to share an intention to resort to individualism against the group mind. A straightforward, non-individualistic concept of shared intentionality is not bound to end up in a collectivist conception of the mind that violates our basic assumption of intentional autonomy. Rather, the specter of the group mind is itself an individualistic artefact that arises from a deep-seated 'Cartesian' preconception concerning the mind, which we should leave behind. This chapter analyzes what the 'Cartesian Brainwash' consists of, and how it is to be overcome by combining the relative strengths of Searle's and Bratman's conceptions.

The third chapter approaches a straightforward reading of shared intentionality by challenging yet another implication of most of the received views of collective

intentionality. These conceptions of collective intentionality claim that in order to share an intention to x (and thus to be a member of the group that intends x), each of us has to intend to do his or her part of x. For the purposes of this chapter, I call this the *Participation Theory of Team Agency*. Against this, Annette Baier has pointed out that there is a *dissident use of "we"*, where the speaker does not take part in the joint activity of the team whose member she still takes herself to be. With his theory of *non-operative membership*, Raimo Tuomela has gone to great lengths to integrate some such cases (i.e. cases of apparently non-participatory membership) into a theoretical framework that in its basic traits seems to conform to the participation theory of team agency. In Tuomela's view, non-operative members have to express some pro-attitude towards the shared goal in order to maintain their status as members.

In this chapter, I cast doubt on the central assumption of the participatory theory of team agency. My argument shall draw on an insight first put forward by Wilfrid Sellars, i.e. the fact that the relation between collective intentions and individual contributions is basically a *normative* relation. If we intend to do x, this provides me with a *reason* to form an intention to do my share in x (this seems true even if we accept John Broome's view that intentions do not provide reasons for the special case of individual intentions). I *ought* to do my part, as it were. As I shall argue, this rules out a strictly participatory understanding of collective intentionality and team agency of a certain kind. At the same time, however, the normative character of collective intentions imposes tight restrictions on dissident uses of "we", some of which are examined in this chapter.

The fourth chapter, which concludes the first part of this book, widens the focus of the discussion. Over the first 2 decades of its history, the analysis of collective intentionality has almost exclusively been concerned with joint intentions, and, to a lesser degree, with shared beliefs. It is only recently that *shared emotions* have started to receive any attention at all. In this central chapter of the book, I argue that, while the received accounts of collective intentionality are important and helpful for an understanding of the cognitive (theoretical) and conative (practical) aspects of shared emotional states, a full understanding of *collective affective intentionality* requires us to focus on the *phenomenological aspects* of emotions, i.e. on *feelings.* An analysis of collective affective intentionality should start out with an examination of what it means to *share a feeling.* The very notion of sharing a feeling, however, seems to be highly problematic, because it is a deep-seated notion that feelings are, by conceptual necessity, *individual.*

This chapter examines three different versions of individualism about feelings, and explores the conceptual room left for a straightforward notion of shared feelings. I claim that, contrary to what is generally thought, there is a sense in which (token) feelings can indeed be shared in the simplest straightforward sense of the word, without violating the plausible assumptions that lie behind individualism about feelings, especially the idea of separateness of persons, and the basic asymmetry between the first and third person in the ascription of feelings. Thus I argue that the fact that some intentional states involve qualitative or phenomenal elements does not speak for an individualistic reading of collective intentionality. The straightforward

conception of collective intentionality sketched in the first part of this book can be extended to the analysis of shared feelings.

The fifth chapter opens the second part of the book, which is devoted to the application of the views developed in the first part to central issues in current social theory and philosophy of social science. The first chapter in this part addresses one of the most important developments in current social science in general and economics in particular, i.e. the rise of *experimental economics.* More specifically, I challenge a particularly successful theorem in the economic theory of cooperation, i.e. *strong reciprocity*. In large part, the idea relies on the interpretation of experimental findings concerning the propensity of human agents for "altruistic punishment".

I argue that the labels used by proponents of strong reciprocity in the interpretation of their data are deeply flawed by an individualistic preconception of human agency. In particular, I argue that (a) altruism is an inadequate label for human cooperative behavior, and (b) an adequate account of cooperation has to depart from the standard economic model of human behavior by taking note of the agents' capacity to see themselves and act as team-members. Contrary to what the proponents of *strong reciprocity* (Ernst Fehr et al.) seem to think, the main problem of the conceptual limitations of the standard model of behavior is not so much the assumption of *selfishness*, but rather the atomistic conception of the individual. A much-neglected question in the theory of cooperation is how the agent's social identity is determined, i.e. how individuals come to think of themselves and to act as members of a group. Taking as an example one of Fehr et al.'s third party punishment experiments, I shall argue that the determination of the agents' identities (and thus the result of the experiment) is heavily influenced by the way the experiment is presented to the test subjects, especially by the collectivity-related vocabulary used in the instructions given to those subjects.[3]

The next chapter (6) is again mostly focused on economic theory. It takes up a question that has stirred considerable debate in recent decision theory and philosophy of economics. The point of departure is this: it seems reasonable to assume that in situations where it is common knowledge that there are multiple coordination equilibria, one of which is the best equilibrium for all participants, and where there is common knowledge of this fact as well as of the rationality of the participating individuals, these agents will choose the strategy that has the best equilibrium among

[3] In this chapter, I use the label "nostrism" to refer to the group-related self-conception of the agent. This label seems a particularly attractive way to make clear that the conceptual dualism of egoism and altruism is not sufficient to capture the structure of human motivation that is responsible for cooperative behavior. The term is used in much the sense in which it was introduced in José Ortega y Gasset's *hombre y el gente*. It was only after I had completed this chapter that I became aware of the fact that the term had been introduced well before Ortega y Gasset. Walther Pembaur used it as a title for the third part of his *Nationalismus und Ethik* (Vienna 1935, pp. 97–115). As opposed to Ortega y Gasset (who would hardly have used the label if he had known about Pembaur's book), Pembaur's concern is not with the structure of action, but with normative ethics. In spite of both this difference and the fact that Ortega y Gasset re-invented the label, and in spite of the fact that Pembaur's political views seem to be at odds with much of what he says in his book (e.g. his explicit rejection of "chauvinism"), Pembaur's national socialism discredits the term so thoroughly that I have decided not to use it in my future writing.

its possible outcomes (similar arguments can be made with regard to the role of salience and focal points in pure coordination games). Yet there is a genuine philosophical scandal here: as has been convincingly shown by authors such as Michael Bacharach and Robert Sugden, our strong pre-theoretic assumption concerning the rationality of coordination is at odds with the most widely held conception of practical reason, i.e. orthodox rational choice theory. There are several reactions to this scandal in the received literature. The first is the most frequent yet least successful one: to deny the incompatibility between standard accounts and common sense. The second is to blame common sense, and the third to blame the standard theory.

I argue in this chapter that we should follow the third path (i.e., that we should hold on to our intuition that choosing the strategy with the best outcome is rational). If this is true, we should look out for a theory that is able to accommodate our pre-theoretic intuition concerning the rationality of cooperation within a plausible overall conception of rationality. In this chapter, I examine some proposals to use the theory of collective intentionality for this purpose. I argue that a strong conception of collective intentionality (such as developed in Part I of this book) could be helpful for this task.

Chapter 7 is aimed at showing how the views developed in the previous chapter fit into one of the most piercing and powerful critiques of Rational Choice Theory that are available in the received literature, namely the one put forward by Amartya K. Sen. In papers such as "Rational Fools" (1977), Sen's central claim is that Rational Choice Theory cannot account for what he calls "committed action", and that committed action plays an important role in human interaction. According to Sen, one of the decisive differences between rational choices and committed actions is that committed agents do not (or do not exclusively) pursue *their own goals*. As Sen's critics have repeatedly pointed out, this claim appears to be nonsensical since even altruistic agents cannot pursue other people's goals without making them their own. Contrary to Sen's claim, it seems that self-goal choice is constitutive of any kind of agency. In this chapter, I aim to show how a strong conception of collective intentionality can be used to *defend Sen's critical claim*. I argue that the objection raised against Sen's critical claim holds only with respect to *individual* goals. Not all goals, however, are individual goals; there are *shared* goals, too. Shared goals are irreducible to individual goals, as the argument from we-derivativeness (as developed in Chapter 2) and the argument from normativity (as developed in Chapter 3) show. It is further claimed in this chapter that an adequate account of committed action defies both internalism and externalism about practical reason.

The last chapter of this second part defines and discusses a certain type of altruistic behavior. The phenomenon is approached from what I propose to call the "Paradox of Altruistic Action". Some philosophers have claimed that there is a contradiction built into the very notion of altruistic action: our standard theory of action implies some degree of selfishness, which is at odds with our intuitive concept of genuine altruism. As there clearly *are* altruistic actions, however, there must be something wrong with this way of putting things. In the usual view, a somewhat *less demanding concept of altruism* is taken to solve the paradox. This solution,

I argue, works well with regard to paradigmatic cases of altruistic behavior. It does not work nearly as well, however, with regard to another class of altruistic behavior, which will be in the focus of this paper. I suggest that, for this other class, we might have to find a different solution to the paradox of altruism. I will make a tentative proposal to that effect. The proposed solution concerns the other side of the paradox, i.e. the standard theory of action. It is suggested that we need to alter or refine some basic assumptions concerning the structure of action in order to accommodate the kind of behavior in question. Here the conceptual tools developed in the first chapter of this book come in handy. I argue that the apparent conflict between the intentional structure of the type of altruistic behavior analyzed in this chapter and the standard notion of action stems from a confusion between individual intentional autonomy and individual intentional autarky as defined and examined in Chapter 1.

In the third part of this volume,[4] the perspective is broadened in its temporal dimension by engaging four selected "classics" in social theory and social philosophy, namely Martin Heidegger, Moritz Lazarus, Gabriel Tarde, and Max Weber. In each of these critical encounters, the aim is twofold. First, it is shown how these authors can still serve as a source of inspiration for our current thinking on the basic conceptual structure of the social in general and on the structure of collective intentionality in particular. Second, it is shown how current theory of collective intentionality, and especially the account of collective intentionality delineated in this volume, can help us to resolve some of the conceptual impasses we encounter in these authors.

The first chapter of this concluding part starts out with an analysis of what has often been perceived as the main problem in Martin Heidegger's *Being and Time.* Heidegger uses the term *Dasein* for the kind of beings we are. His description of everyday *Dasein* distinguishes between the sphere of more or less solitary instrumental or goal-oriented action, and the social sphere qua realm of convention and communication. In Heidegger's analysis, there is a tendency to associate the public sphere with what he calls *inauthenticity*. A widely perceived and often criticized consequence of this is a rather solipsistic conception of *Dasein*'s authenticity, and it seems obvious that something has to be done about this if we are to take Heidegger's analysis of *Dasein* seriously.

Some American Heideggerians have recently ventured into revising this conception by showing that the sphere of instrumental or goal-oriented action is itself constituted by public norms and conventions, so that there is no outside to the public sphere. As much as I agree with the basic thrust of these authors, I argue that this reading simply means pulling the teeth from Heidegger's notion of inauthenticity, leading into a conventionalist view of *Dasein*. Defending Heidegger's anti-conventionalism against his current pragmatist interpreters, I shall propose

[4] It was only after I finished working on this book that I noticed that all chapters on the role of collective intentionality in current social science (Part II) are focused on economic theory, while the chapters dealing with historical interpretations in part III are largely devoted to sociological theory. This does not mean that I endorse the view that is sometimes voiced by economists according to which economic theory has succeeded sociological theory as the leading strand of theory in social science. In large part, this choice of focus is due to biographical contingencies.

another solution to the problem instead. Not surprisingly, the solution I propose uses the concept of collective intentionality. According to my suggestion, the problem is not the conceptual distinction between goal- and norm-oriented action, as the pragmatist interpreters assume, but rather the way in which Heidegger conceives of goal-oriented action. Heidegger simply neglects the dimension of *cooperation* in his depiction of skillful coping. In fact, a closer look at some of Heidegger's lectures from the time immediately following the publication of *Being and Time* shows that at this decisive point on his "*Denkweg*," Heidegger himself briefly ventured deep into a strong conception of collective intentionality, but ultimately shied away from drawing the full consequences and so ended up with the infamous collectivistic conception of the "*Dasein* of the people" in the early 1930s.

I argue that an adequate understanding of the social dimension of *Dasein*'s authenticity should take up this line of thought where Heidegger left off, and that this dovetails with a straightforward conception of collective intentionality as developed in Part I of this volume. I furthermore argue that there is a basic fact about collective intentionality that can be learned from Heidegger, which concerns the role of reflective attitudes from the side of the participants. Here, I also take up the issue of collective identity as touched on in Chapters 5 and 6.

The next chapter picks up the question of reflexive attitudes, and connects this issue with the problem of the collective mind as presented in Chapter 2. Since the end of World War II, the only role left for the members of the extended semantical family of the collective mind in social theory and social science has been that of a specter that needs to be exorcised by means of a firm commitment to methodological or ontological individualism. Individuals and their mutual relations, rather than "spooky holistic entities", are seen as the proper object of social science. The only appeal made to such entities is when it comes to justifying the individualistic setting of current social theory (for a series of good examples for this use cf. Chapter 2).

Yet there is another side to the coin. In recent years, some new overtones can be heard in the discussion about social ontology and the philosophy of social science. In the current struggle for a more adequate understanding of the structure of the social world, there are some ideas floating around that, at first sight at least, seem to bear a striking resemblance to other members of the wider family of the collective mind. Leading participants in the contemporary debate such as Margaret Gilbert, Philip Pettit, or Raimo Tuomela use terms like "plural subject" (Gilbert), or "groups with minds of their own" (Pettit), or speak, somewhat more cautiously, of "modern counterparts of group minds" (Tuomela). Many philosophers are interested in forms of collective agency that cannot simply be reduced to the agency of the participating individuals. Some philosophers have even started to openly consider the possibility that there is a sense in which personhood might be attributed to collectives, in the proper sense of the word, that goes beyond the merely formal meaning of corporate personhood in law.

Considering these and other examples, one might ask: is the collective mind really as dead as it seemed? And if it is still alive, or has come back to life in the current debate, what should we make of this fact? In this chapter, I focus on the

most infamous of all conceptions of the collective mind, the *Volksgeist*. If so many other members of the family of the collective mind have already found their way back into the current debate, what should prevent the return of the *Volksgeist*?

I start out with an analysis of what seems to be deeply wrong about the very idea of the *Volksgeist,* and of why the *Volksgeist* has played such a fatal role in the history of political thought. I then turn to a conception of the *Volksgeist* that, at first sight at least, seems to be free of these flaws, namely that of the German social psychologist Moritz Lazarus (1824–1903). It is not without reason that, after long years of near-total oblivion, a selection of Lazarus' works has recently been published, for upon closer inspection, Lazarus' concept of the *Volksgeist* reveals surprising similarities to current conceptions of social identity. I argue, however, that this resemblance should not be taken as an argument for Lazarus' conception, but rather as an argument against our current mainstream thinking about social identity. I show how some of the flaws of the very idea of the *Volksgeist* permeate Lazarus' conception, and how this affects our thinking about collective identity. Here, Heidegger's insight into the pre-reflexive character of intentional togetherness as developed in the previous chapter is important.

The next chapter addresses an altogether different issue. *Meme theory* confronts us with a rather unflattering image of ourselves. In Daniel C. Dennett's words, conscious selves are nothing but the 'vehicles' or 'nests' of the true heroes of the evolutionary story of culture, memes. In the memetic view, cultural evolution is not about 'us', but about 'them': such units of culture as those mentioned by Richard Dawkins: "tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches". In this chapter, I take a critical look at some of the assumptions on which this memetic 'shift of perspective' is based – assumptions that turn out to be highly problematic indeed. In the first step, the image of the self as a 'meme nest' is traced back to its neo-Darwinian origins. Meme theory is built directly on the model of genetic evolution, and genes are understood in terms of DNA sequences. As some considerations concerning the ontology of memes (which I shall present in the second step) reveal, there are fundamental differences between this view of genes and that of memes which cannot be accounted for within the memetic view. In the third step, the French sociologist Gabriel Tarde's (1843–1904) idea of 'evolution by association' is introduced as a convincing alternative to the memetic idea of cultural evolution. Writing almost a century before the term 'meme' was coined, Tarde put forward a theory that already contained those insights that make memetics seem attractive, without falling into the mistakes so pervasive in current memetic thinking. Tarde was safe from the fatal tendency to model cultural evolution on the mistaken model of genetic evolution that has all but discredited memetics. In a "Tardean" view, our role in cultural evolution is neither that of a completely sovereign subject, nor that of a mere meme vehicle. Rather, the emerging image corresponds to the view of embedded agency as developed in the theory of plural action.

The concluding chapter in this volume takes up the issue analyzed in Chapter 3, i.e. the role of normativity in joint intentional activity. The chapter comes in three sections. The first section discusses Max Weber's (1864–1920) concept of

consensus. Weber's use of the term is seen as an attempt to accommodate social normativity within his general instrumentalist framework of social action. The second section shows why Weber failed, and how his failure prompted the two leading strands in German social theory to depart from Weber's intentionalist action theoretical framework. Niklas Luhmann's systems theory, as well as Jürgen Habermas' theory of communicative action, can be interpreted as attempts to find a theoretical framework that is better suited to the task of integrating social normativity into the theory of joint intentional activity. In the third section of this chapter, I argue that neither of these alternatives are convincing, and that a straightforward theory of collective intentionality and plural action is needed in order to develop an adequate view of the connection between social normativity and instrumental action.

Needless to say, no part of this volume can make any claim to completeness. My aim at this stage in the development of the analysis of collective intentionality is to strengthen the case for an intentionalist approach to the basic issues in social theory and social ontology by addressing core controversies and pointing out connections rather than by providing a water-tight analysis. My hope for this volume is that it will help both to deepen and broaden our perspective on collective intentionality, and to draw more attention to this novel approach to the structure of the social world.

Part I
Collective Intentionality Reconsidered

Chapter 1
Plural Action

Concepts and Problems

I call *plural* those actions that require the participation of at least two individuals (sociality condition) acting in pursuit of one and the same goal (plurality condition). Examples of plural actions are: going for a walk together, jointly writing a paper, or playing a symphony. Even though plural actions abound in our lives, they have been somewhat neglected in philosophical analysis. Part of the reason for this is that plural actions do not seem to fit easily into our standard philosophical conception of agency. Whereas any singular action can be attributed to a single individual agent – the only kind of agent standard theory of action knows of – plural actions seem to require a different kind of agent (the plural agent problem). In the first section of this chapter, I shall use the intuitive idea that one cannot intend what one takes oneself to be unable to perform to approach the plural agent problem, and situate plural actions within a taxonomy of action types (§1). I then turn from action theory to common sense. In contrast to action theory, common sense seems to have no difficulty whatsoever in coming up with suitable agents for plural actions. There are at least three different common sense solutions to the plural agent problem: plural actions are either attributed to *collective agents* (such as in the case of Parliament's passing a law), to *powerful individuals* (such as in the case of Caesar's defeating the Helvetii), or to a plurality of individuals jointly intending an action (such as in the case of a bunch of friends going for a walk together). These three replies correspond to three different models (or perhaps types) of plural agency. I propose to call them the collective agent model, the influence model and the teamwork model, respectively, and I shall argue that the teamwork model is the most fundamental of these.

I shall then turn to a somewhat more detailed discussion of each of these models, and examine the reasons why they are met with so much reservation (or even resistance) from the side of action theory. As to the *collective agent view* (§2), it is obvious that many authors are still reluctant to ascribe intentions and actions to collectives. This reluctance found its classical expression in methodological individualism of the Weberian kind. I examine a reason Max Weber might have had for not admitting collective agents to the basic level of intentional interpretation, and I conclude that he believed collective agency to be incompatible with what I suggest to call *individual intentional autonomy*. I propose this label for the view that each individual is an *agent* of his or her own, i.e. that his or her behavior should

H.B. Schmid, *Plural Action: Essays in Philosophy and Social Science*, Contributions to Phenomenology 58, © Springer Science+Business Media B.V. 2009

be interpreted as *his or her own actions*. Looking at current conceptions of collective agency, I argue that, contrary to what Weber seems to have thought, collective agency is compatible with individual intentional autonomy, so there is no reason not to accommodate a robust conception of collective agency in action theory.

The following section (§3) examines the *power or influence model* of plural agency, according to which the leaders and authority figures have a claim on the ownership of a plural action. Its main intuitive problem seems to be that, by ascribing the plural action in question to one single individual, it bypasses the other participants' own individual agency. According to this view, all participating individuals have a claim on their own individual *contribution* to the plural action, and no individual has a claim on more than that, so that the plural action as a whole cannot be attributed to one individual, however powerful he or she might be. It might seem that this view is a direct consequence of the *intentional autonomy* of the participating individuals, but I shall argue that this is mistaken, and that there is a *further* and *more problematic claim* involved in this view, which I shall call *individual motivational autarky*.

The assumption of motivational autarky is that each individual acts on a motivational agenda of his or her own, i.e. that the interpretation of each individual's behavior has to bottom out in *his or her own volitions or pro-attitudes* (rather than in some other individual's volitions or pro-attitudes). I argue that intentional autonomy does not imply motivational autarky, and that "non-autarkical" behavior might actually play an important role in many cases of plural action (this issue is further pursued in Chapter 8 below). I conclude this section with a discussion of why the autonomy assumption and the autarky assumption have always been mixed up. My thesis is that while intentional autonomy is an universal feature of human agency, and indeed an essential feature of what it means to be an agent, motivational autarky is something very different: a very strong *normative ideal* in our particular culture.

Turning finally to the *most basic form* of plural action in §4, I argue that the main problem *teamwork models of plural agency* have to cope with is that of reconciling the unity of action with the plurality of agents. This is precisely what, in the current literature, the concept of *collective intentionality* is invoked for: many individuals can intend and claim shared ownership of a plural action insofar as they share the respective intention. Most philosophers of collective intentionality, however, are very reluctant to admit a *straightforward understanding* of the sharedness of intentional states. The predominant view is a distributive reading of collective intentionality, according to which individuals cannot *literally* share an intention, and that each individual has his or her own intention when they intend to do something together.

Behind this view lies what I propose to label *intentional individualism:* the view that all intentionality is *some individual's*. I shall argue that this *distributive* reading leads to circular conceptions of collective intentionality. I believe that the reason why most philosophers of collective intentionality endorse a distributive conception is that they believe that intentional individualism is implied in individual intentional autonomy. I shall argue that it is not, and that the theory of plural agency will make great headway by dropping intentional individualism, and endorsing a stronger conception of *intentional commonality*.

The concluding §5 wraps up the line of argument developed in this chapter, and closes with an observation on the occasion of the first centenary (2008) of the term "methodological individualism".

§1 The Plural Agent Problem

In the earliest stages of the Apollo Program, John F. Kennedy once went to Florida to visit Cape Canaveral. On his tour through the facilities, he asked a technician what his task was. The anecdote has the man giving the following answer: "To land a man on the moon and returning him safely to the earth before this decade is out, Mr. President!"

What is so unusual here that people bother to retell the story? The clue, it seems, is the huge gap between what the worker claims to be his task, on the one hand, and what he is actually able to do on the other. If the man is really *serious* about what he claims to be his task, sending a man to the moon is what he in fact *intends* to do; if this is the case, however, he expects *far too much* of himself. Sending a man on the moon simply exceeds a single technician's possibilities.

Thus this anecdote sheds some light on how we *normally* think about the relation between intentions and abilities. What seems to be at stake here is something we might call the Principle of *Intentional Self-Confidence*. This principle puts some constraints on the range of things one can intend to do. It states that an agent's intentions must be *in tune* with what he or she takes to be *possible*, given her abilities and the opportunities at hand.[1] In the briefest (negative) version, the principle reads as follows:

(A) The Principle of Intentional Self-Confidence: One cannot intend to do what one takes oneself to be unable to carry out.[2]

This needs some explanation. First, the principle of intentional self-confidence is perfectly compatible with the fact that "intend" does not imply "can". People might well intend the *objectively impossible* just as long as they don't *take it* to be impossible. Also, the principle does not rule out the possibility of certain forms of aiming at the *subjectively impossible*.[3] And naturally, the principle does not entail

[1] For an analysis of the concept of ability cf. Kenny 1976. Useful ideas – especially the distinction between the subjective and the objective components of ability – can be found in Löwenstein 1911.

[2] For an early version of this principle cf. Baier (1970). The main difference between the usual way of putting the limitation claim and mine is this. Taking oneself to be able to do x is usually assumed to be a matter of *belief*, i.e. a cognitive intentional state. Self-confidence, by contrast, is an *emotion*, i.e. an affective intentional state.

[3] In his "Impossible Doings" (1992), as well as in some later papers on the topic, Kirk Ludwig contested that claim. Ludwig discusses the following example. P assumes (with certainty) that the battery of his car is dead. Upon another person's request, he turns the ignition key. Contrary to what he expects, the engine starts. Ludwig claims that it would be wrong to say that P started the engine unintentionally. I agree with Ludwig that there are some cases of trying the subjectively impossible

the claim that, in order to intend to do A, one has to believe that one *will* A.[4] In his *change in view*, Gilbert Harman (1986) discusses the example of a sniper who takes himself to be a terrible marksman, but still intends to kill the ambassador, in spite of his self-doubts. This is not in conflict with the principle. Sometimes, agents have high intentional *ambitions*, and seriously intend to do things they perfectly well *know* they might be unable to achieve after all (indeed, such ambitious intentions are common at the start of any project, such as the intention to write a book on plural agency). In other words, one does not have to be *certain* to be able to perform what one intends to do. Life would be boring indeed if we limited our intentions to things we perfectly well know we can do. This is to say that intentional self-confidence might well be *minimal*. Even the faintest hope of achievement is enough. All the principle states is that intentional self-confidence cannot be *zero*, for one cannot intend to do what one is perfectly *sure* of being unable to perform. If I *know* that the restaurant opens only at 6 p.m., I cannot *intend* to have lunch there at noon. If I still choose to go there, and if this isn't a case of conflict between what I know my abilities to be in my head and what I feel able to do in my guts, my intention is a different one: to act *as if* I didn't know the restaurant's opening hours, to knock on the closed door as if in surprise, or any other aim along these lines.

I have already emphasized that the principle of intentional self-confidence is not in conflict with the possibility that one might be *mistaken* in what one takes oneself to be able to perform. One can always *misjudge* one's forces and abilities, and expect too much or too little of oneself. This is perfectly possible, but to the degree that this happens *systematically* and *under normal circumstances*, it renders intentional self-confidence *irrational*. Intentional Self-Confidence is *rational* to the extent that under normal circumstances one's intentions are *in balance* with one's *actual* forces and abilities.

Let's call this the *rationality specification* of the principle. It allows for two directions of imbalance: one can either overrate or underrate one's forces and abilities. In other words, intentional self-confidence can be irrational in two ways – for lack of better terms, let's simply label them "objective" and "subjective", respectively. Intentional self-confidence is *objectively irrational* if one intends to do a thing which one is *generally* and *under normal circumstances* incapable of performing. (It is always possible to fail at a task one takes oneself to be perfectly capable of performing. If this happens *by chance* or *due to unusual and unforeseeable circumstances*, this does not render one's intentional self-confidence irrational.) Conversely, intentional self-confidence is *subjectively irrational* if one fails to form an *intention* to do something one wishes to be done *for the sole reason* that one takes oneself to be

---

where "trying" does not function as a proper action term. I argue, however, that in such cases, the agent must take himself to have a chance at success, however minimal, which might be in conflict with his conscious assessment of the situation. Sometimes an agent's intentional self-confidence is not in tune with his or her beliefs concerning his or her ability. If this is true, Ludwig's point does not prove that it is not the case that people cannot intend to do what they take themselves to be unable to do. The question is how to understand the "taking": insofar as it is belief, Ludwig is right; insofar as it is self-confidence, he is not.

[4] This claim is often ascribed to Paul Grice (1971) and J. David Velleman (1989).

unable to do it, when it is actually well within one's forces and abilities. This second form of irrationality consists in an *understatement* of one's forces and abilities.

Again: neither of these two kinds of irrationality is incompatible with the principle as such. The principle of intentional self-confidence is a *conceptual* principle. It is part of how we use the term "to intend".[5] As such, however, the principle does not say that intentional self-confidence is always *rational*. Rationality, in other words, is a *normative standard* for intentional self-confidence, not a *conceptual requirement*, as those pervasive cases of both varieties of *irrational* intentional self-confidence show rather clearly.

The fact that not all intentional self-confidence is rational seems to be precisely what makes the NASA technician's reply in the above anecdote funny. The man's reply takes intentional self-confidence to its *objectively irrational extreme*. By taking his task to be to send a man to the moon, he takes himself to *be able* to do such a thing, which he is clearly not, because what single workers can do is limited to such things as wiping factory floors, assembling parts of rocket stages, etc. Thus the worker grossly and grotesquely overestimates his forces.

I will not delve any deeper into an analysis of intentional self-confidence here, but rather use the principle as a guide to a quite different issue at stake in the anecdote. If sending a man to the moon couldn't rationally be *a single worker's intention*, because what single workers can do is limited to much more moderate tasks such as assembling rocket parts, the question arises: *whose task was it, then?* Who could *rationally* ever be so *self-confident* as to intend to land a man on the moon? This is the question I wish to address in the following. Before looking at possible candidates for this role, however, I should first make sure that this question does indeed make sense. It does so only if intention is conceived of in action-referential terms, which some authors claim is not necessary. And it does so only if something like the moon expedition can be described as *one* action. Is this true, and if so: to what particular *kind* of action do such things belong?

First a remark on the question of whether or not intention should be conceived of in action-referential terms. Intention is action-referential insofar as it is an intention *to* A. This seems the most natural way of putting intention, but this places tight restrictions on the possible objects of intention. A needs to be an action, and as the only actions one can intend *directly* seem to be one's own, it appears that the only objects of intention are *one's own actions*. In action theory, there is a tendency to claim that possible objects of intention extend beyond one's own actions (cf., e.g., Bratman 1987; Vermazen 1993). These authors claim that the objects of intentions are *propositions* rather than actions. Thus intention should be conceived of in *propositional* rather in *action-referential* terms. In their view, intention should be seen as intention *that p*, where p can be any state of affairs, and even an action whose subject is not the subject of the intention. This considerably widens the scope of intention. The question is: should the technician in the above anecdote have said "I intend *that* a man be landed on the moon and returned safely to earth"?

[5] Needless to say, the constraints articulated in the principle of intentional self-confidence apply only to *intentions*, not to wishes and other intentional states.

Wilfrid Sellars (1992: 183ff.) has claimed that propositional intentions are expressions of practical commitments only by virtue of their conceptual tie to action referential intentions. Sellars says that the intention that X, when made explicit, is the *intention to do whatever is necessary to make it the case that X*, which is action self-referential, because again, one has to perform the doing oneself. I think he is right. But it is completely sufficient for present purposes to accept action referential intention as *one important kind* of intention, and that the action-referential mode of expressing intention should not be abandoned completely, even though it might not be the only way of thinking about intention.

What about the second question concerning the unity of action? Let me start by mentioning four fairly uncontroversial features of the concept of action. First, for there to be an action there has to be some kind of agent, i.e. somebody to whom the action is attributed, and who can be held responsible for its consequences according to our normative practices. Second, action requires some kind of *behavior*, of which the agent is in a certain degree of control (typically consisting of the agent's own bodily movements). Third, some *goal* is needed, i.e. something the agent *wants*, a state of affairs towards which the agent has some kind of *pro-attitude* or "desire" in the wide sense of the word. In the context of action, goals are conceptually tied to *intentions*. If a complex of behavior is taken to be an action, it is assumed that the agent is in fact *trying to achieve* his or her goal, i.e. that the goal is the condition of satisfaction of an intention. And fourth, there has to be some connection between the agent's goal and the complex of behavior in question. The behavior has to be *minimally rational*, i.e. the agent has to show at least some minimal degree of *concern* about the behavior's being suited as a means to the end (however successful or unsuccessful she might be at this task).

It goes without saying that although these characteristics may be necessary conditions for actions, they are certainly not sufficient ones. We do not have to delve any deeper into action theory here, however, in order to answer the simple question: was man's travelling to the moon an action? If we leave aside for the moment the open question concerning the agent, it seems that the moon expedition meets all conditions. As to the goal-directedness and rationality of behavior, it seems to even be a *paradigmatic* case of an action. There clearly was a goal, and not only was the goal obviously *intended*, but the achievement of the goal was also permanently monitored, with a constant effort to choose suitable means to the end. Thus it seems clear: the moon expedition *was* an action – *if* the one open question can be answered: *if* an agent can be identified, i.e. *if* it is possible to answer the question *whose* action it was.

Before we come to that, let's just assume for the moment that a plausible answer to the open question *had* been given, and have a closer look at the *type* of action to which something like the moon expedition would belong. I propose the term *plural action* for the kind of action in question. Plural actions are *social actions*. I call social those actions that require the participation of more than one individual (I label this the *sociality condition*). Let's call the class of non-social actions *solitary* actions (it is still controversial whether or not this class contains any elements, i.e. whether or not a hypothetical lifetime Robinson Crusoe could be an agent). There

are two ways in which sociality might be required: either *logically* (as in the classic case of marrying, which as a matter of conceptual necessity takes at least two), or *contingently* (as in the case of building a house, which one could do all by oneself if only one was a little stronger). Accordingly, there are two kinds of social actions. Sending a man to the moon belongs to the class of contingently social actions. I'm not concerned with *this* distinction here, however, but with the one between *singular* and *plural* actions, which is independent of the distinction between logically and contingently social actions.

The distinction between singular and plural actions is a matter of the *goals* pursued in each case. Singular actions are social actions in which the participating individuals pursue *different goals*. By way of example, consider the case of my taking a plane back home for Christmas, which presupposes, among other things, some mechanic's activity, but my spending Christmas at home and his getting the jet engines going are quite different goals. In contrast to examples such as this, *plural actions* are social actions in which the individuals pursue the *same goal*.

To give an example: our playing a duet together requires that we have this goal in common. If you simply aim at performing your part (as long as I perform mine), and *vice versa* – if our goals are, in other words, *different* goals – we may make our way through the score, but we certainly won't be playing a duet. If our playing is to be a duet, our goal has to be the *same*. I call this the *plurality condition*. If Saturn V had been produced with the sole aim of selling it to the highest bidder, the production of Saturn V would not have been part of the plural action of sending a man to the moon. To be as explicit on this point as possible: plural actions require more than the participants' having *similar* goals. Just because each individual in a group has a similar or the same *type* of goal, or even goals with the same *content*, does not make the activity in question a plural action. The goal must ultimately be *one and the same* goal. In brief, the main characteristics of plural actions are these: *many* participants, *one* goal (Fig. 1.1).

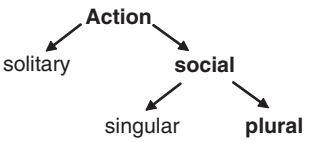

Fig. 1.1 Taxonomy of action types

Plural actions have long been rather shamefully neglected in action theory. Only in the course of the last 2 decades has the phenomenon started to attract any attention. In the meantime, however, a small, but rapidly growing debate on the structure of plural action has developed. It is characteristic of much of this debate that *small-scale examples* are used to discuss the structure of plural action. Activities such as going for a walk together (Gilbert 1996), jointly operating a water pump (Bratman 1999), preparing a Sauce Hollandaise by one pouring the oil and one stirring the sauce (Searle 1990), or pushing a broken-down car together (Tuomela 1995) serve as illustrations of the phenomenon. By contrast to this, my choice in this chapter is a large-scale example; apart from wars and military expeditions, the

Apollo program was probably among the most extended plural actions in the entire history of mankind (I will turn to smallest-scale examples below).

I have chosen this example because the large scale helps to illustrate what I see as the *crucial problem in the theory of plural action*. It is this: If the principle of intentional self-confidence is valid, plural actions require a particular *kind of agent*: one that can *rationally* take him- or herself to be able to perform an action which can only be carried out with the joint efforts of many? But *who can possibly fit that bill? What are plural agents?* Are there any plausible candidates that conform to the principle of intentional self-confidence *objectively rationally* (in the sense defined above) with regard to *plural actions?*

Common sense offers no less than three types of candidates: in everyday parlance, we routinely ascribe plural actions to *collective agents*, to *influential individuals* (or leaders), and to *jointly intending individuals*. These three types correspond to three commonsensical models of plural agency: the collective agent model, the influence model, and the teamwork model. In the following sections I will characterize each of these models in turn, and make some comments on why these intuitive notions have not usually been well received in action theory (to say the least), and on the main obstacle standing in the way of a deeper and more adequate analysis of plural agency.

§2 Collective Agents and Individual Autonomy

According to what I propose to call the collective agent view, plural actions may be intended (and indeed performed) by *collectives*. The plausibility of this view can easily be illustrated with our example. Just remember Kennedy's choice of words when he announced the start of the Apollo program in May 1961: "I believe that *this nation* should *commit itself to achieving the goal*, before this decade is out, of landing a man on the moon." So maybe the nation really did it after all, or perhaps NASA did it. Another, somewhat less plausible candidate for that role would be *mankind* – the collective Neil Armstrong invoked when he was setting his foot on the moon. The advantage of this model is that it *avoids* the problem of rational intentional self-confidence by invoking an agent that matches the size of the task in question. The agent is not an individual agent, but a super-agent over and above the heads of the participating individuals. Such agents, it seems, need not worry about expecting too much of themselves when they form an intention to carry out a plural action. Given their size, they can be rationally self-confident in their ability to do such things. The only question is: do such agents really exist?

Most prominently, the forefathers and founding fathers of *methodological individualism* rejected the assumption of collective agents. Thus Max Weber famously stated that the only agents which social science recognizes are individuals.[6]

[6] Weber articulates this central precept of his methodology in the following way. When discussing social phenomena, we often talk about various "social collectivities, such as states, associations, business corporations, foundations, as if they were individual persons" (Weber [1921] 1980: 13).

What is the reason for Weber's view? It has repeatedly been pointed out that Weber's rejection of collective agency should be seen in the context of his commitment to the method of *intentional interpretation* (Heath 2005). And this is indeed what Weber says: *only individuals* "can be treated as agents in a course of subjectively understandable action" (Weber [1921] 1980: 6). Collectives, Weber seems to think, are simply *not suited* as objects for intentional interpretation. Since, in everyday life, we often seem to have no difficulty whatsoever in ascribing intentions and actions to collectives, however, one might wonder what reason Weber might have had for this claim.

Unfortunately, Weber does not expand on this any further. But one can think of a whole series of arguments for this view. First, it is well known that Weber defines action by meaning (*Sinn*) and behavior, both of which – in Weber's view at least – seem to be essentially individual. Another reason might be that, within the framework of Weber's analysis, action is supposed to play the role of the *explanans*, with the collectivity being the *explananda*[7] – this naturally excludes plural action from the class of *explanantia*, because this kind of action seems to be the sort that already involves collectivity (cf. the *plurality condition* mentioned above).[8]

I certainly do not underestimate the role of any of these arguments for Weber, but I think that his *basic concern* is yet another one. The worry is this. If we were to treat collectives as agents, individual agency would be somehow conceptually *compromised* or *impaired*. The point of departure of *this line* is Weber's firm commitment to the view that individual behavior *is* the proper object of intentional interpretation. Individuals are *agents*. Their behavior instantiates *their actions*. This commitment to the agency of individual persons, Weber seems to think, is *incompatible* with the assumption that there are any agents other than single individuals, and in particular with the assumptions that there are collective agents. For if collectives were proper agents, the participating individuals would be nothing more than the mere *instruments* or executing organs of some collective will, and would not be the proper agents behind their behavior anymore. Thus it seems that, insofar as individuals are to *be treated as agents* in the interpretation of social phenomena, collectives simply cannot be so treated. Admittedly, Weber never explicitly says so, but I believe that

---

While Weber does not take issue with any such everyday talk at all (he even admits that for other epistemic purposes, the assumption of collective agents might indeed be quite "useful"), he stoutly opposes its use in scientific interpretation.

[7] The Weberian project is to explain collectivities as "consequences and organizations of individual actions" (Weber [1921] 1980: 13).

[8] Along this line, Weber's is simply a reductivist view: while it is not necessary to use collectivity concepts to describe individual agency, all collectivity concepts can be translated into descriptions of aggregates of individual actions. The only reason why social science cannot fully *do away* with collectivity concepts on the lowest level is, according to Weber, that collectivities are part of the *content* of individual intentional mental states. People happen to *believe* that there are collectives, and they act on this belief. Insofar as a certain type of individual action is the object of social science, collectivity concepts cannot completely be ignored. But clearly, the order of explanation goes from the individual to the collective. There are collectives, because people *think* there are collectives, and not the other way around.

this is the worry in the back of his mind that leads him to oppose the group agent assumption.[9] Because individuals *are* to be interpreted as agents, there can be *nothing but* individuals on the most basic level of intentional explanation. For any other type of agency would *displace* the participating individual's agency. Even though his conclusion might be controversial (we shall come back to this shortly), I think that Weber's premise, i.e. his firm commitment to the individuals' own agency, is basically right. In the current debate, the commitment to individual agency seems to be so universally accepted that it is not even identified as such. Even those who reject the *conclusions* drawn by Weber, and believe that plural agents are an important feature of the very basic structure of social reality, seem to take it for granted that this is compatible with a robust notion of individual agency. I shall call this commitment, which I believe to be at the heart of methodological individualism, the principle of *individual intentional autonomy*. In its shortest formulation, the assumption is the following:

(B) Individual Intentional Autonomy: Under normal circumstances, each individual's behavior instantiates *his or her own actions*.

"Normal circumstances" exclude such cases as mere reflex behavior, which does not instantiate any action at all. Admittedly, the use of the term "autonomy" is somewhat unusual in this context. In the current debate, autonomy is normally taken to involve such highly complex and elaborate structures as self-transcendence, motivational hierarchies, and reflective self-management (cf. e.g. Bratman 2007: 162ff., 195ff.). None of these is presupposed or involved in what I call individual intentional autonomy, even though I dare to claim that, conversely, intentional autonomy in the sense defined here is one important presupposition of all of these more ambitious and richer philosophical concepts of autonomy. In other words, my use of the term autonomy underlies any of the current controversies revolving around this concept. Intentional autonomy refers to a very basic and elementary way in which individuals are responsible for their behavior as agents, in which their behavior can be ascribed to them *as actions*, and in which they can – to introduce a metaphor which I will use repeatedly below – claim *ownership* of their action.

For further clarification of the term intentional autonomy, I introduce its equally neologistic counterpart, *intentional heteronomy*. An intentionally *heteronomous* individual's behavior, were it to exist – which I doubt – would instantiate none of the respective individual's *own* actions, but rather that of *another agent*. In other words, intentionally heteronomous individuals would have to be taken as behaving on another agent's *remote control*, as it were. They would in fact be what we might call *intentional zombies*, to add yet another sort of zombie to the philosophical literature. In contrast to this, the principle of individual intentional autonomy

[9] It is not easy to feel the threat of collectivism now, and perhaps collectivism was never much more than a specter that haunted this debate, but I assume that, in this role, it has been quite effective. Even Ludwig Gumplowicz (1928), who to my knowledge went farthest among early social scientist in asserting the independence of the intentionality and agency of collectives from the intentionality of individuals, asserted that any explanation of social phenomena ultimately bottoms out in motivations for individual actions.

states that individuals are *not* intentional zombies because they do *not* behave on remote control. I assume that this claim is uncontroversial. Intentional zombies abound in philosophical thought experiments (e.g. Mele 2003), in some radical interpretations of the possible effects of hypnotism, in self-reports by schizophrenic patients (Spence 2001; Marcel 2003), and in American Sci-Fi. It seems generally accepted, however, that the everyday social world is not populated by intentional zombies.

So much for the defense of Weber's core claim. The critical question, however, is: is Weber right in assuming that individual intentional autonomy is incompatible with a solid conception of collective agency? I think that he is not, and I take it that this has been sufficiently established by the recently renewed interest in the agency of collectives. There are robust conceptions of collective agency on the market which are perfectly compatible with individual intentional autonomy. This is particularly obvious in Philip Pettit's recent work. In his analysis of the *discursive dilemma*, Pettit has developed the view that (certain types of) collectives can be interpreted as intentional subjects. He even ascribes to them some sort of personhood (Pettit 2002; cf. Rovane 1998), such that, groups can have "a mind of their own" (Pettit 2003).

In his analyses, Pettit's concern is with the *rational unity* of the groups' perspective, which under some circumstances, requires some measure of *discontinuity* with the participating individuals' own perspectives. The phenomenon to which Pettit draws our attention is that the rational unity of a group perspective sometimes requires that this perspective be *distinct* from that of any of the participating individuals'. These collectives are genuine agents. But it becomes more than clear in analyses such as Pettit's that, contrary to what I think Weber's worries were, this does not compromise or displace the agency of the individuals. In Pettit's conception of collective agency, plural subjecthood is solidly grounded in the volitions of the participating individuals. Groups have a sort of agency of their own based on the participating individuals' *insight* into the problems of aggregating individual decisions to collective decisions, and on the participating individuals' *choice* to get their collective act together in avoiding the pitfalls of the *discursive dilemma* and to act consistently and rationally *as a group*. Forming a collective agent does not compromise or displace, but rather *presupposes*, individual intentional autonomy.

Thus, contrary to a worry that still seems to be in the back of the mind of many action theorists, genuine collective agency does not compromise individual agency. There is no reason why action theory should treat conceptions of collective agency with so much reserve. This does not mean, however, that *any* kind of plural agency can be interpreted along the lines of the collective agent model. Collective agency is but one kind of plural agency. The extension of the collective agent model is limited to those cases where the distinction between the collective agent on the one hand and the participating individuals on the other has some intuitive plausibility because this is the way the participating individuals *themselves* interpret their situation. This is particularly true of Hobbesian *personae fictae* – but there is no *persona ficta* involved in smaller cases of plural actions such as going for a walk together. If the two of us go for a walk together, there are only two agents involved in the process – not one, and most certainly not three. Thus the collective

agent model provides a solution to the plural agent problem only in some selected cases. For other plural actions, we seem to need different plural agents. Yet there is an even farther-reaching reason for doubt as to the scope of collective agent-explanations of plural actions. It seems clear that for a collective agent to emerge from the agency of intentionally autonomous individuals, there has to be some sort of *agreement* between those individuals. Such agreements usually presuppose communication. Acts of communication, however, are plural actions. Thus it seems that collective agents presuppose plural agents which are not collective agents. Let us therefore turn to the remaining two common-sense replies to the question of plural agency.

§3 The Dogma of Motivational Autarky

The second type of answer to the "whodunit" question concerning plural agency is this: those individuals who were in control of the project (according to its institutional structure) did it. Let's again take the Apollo program as an example. While the statement "My task is to land a man on the moon" might sound rather silly coming from a simple technician's mouth, it doesn't nearly as much coming from, say, NASA's chief administrator in his leather armchair, or indeed from the President's own lips. It seems that such people's claims to rational self-confidence are simply much better substantiated than those of lower-ranking individuals. To put it as bluntly as possible: great people can do great things. In the memory of Kennedy's recently (2007) deceased court historian, I'm tempted to label this *second* model the *Arthur M. Schlesinger-view* of plural agency, but for the sake of brevity, let's stick to the label *influence model of plural agency*. Admittedly, this is a somewhat patriarchal notion, and, to say the least, it is not very popular in the current humanities and social sciences. Among its advantages, however, is the fact that it is deeply rooted in everyday talk. Behind the erection of the palace of Versailles was Louis XIV intention; it was Vasco da Gama who successfully searched for the sea passage to India, etc. etc. This view models plural agency very closely on the paradigm of singular agency, which makes plural agency look somewhat less unfamiliar. But therein lies the central problem of the model. In attributing plural actions to single individual agents, it makes it look like these leaders had performed their great deeds all by themselves.

Thus the model seems to suggest that leaders have many hands, feet and eyes, not just two of each. The agency of the other individuals involved in the process is simply bypassed, and their individual contribution remains completely unaccounted for in this view. This cannot be right, and it goes against the grain of a deep-seated "democratic" conviction in the theory of social action. This conviction has it that individuals – however powerful they might be – can be attributed only their *individual contribution* to plural actions, and not the plural action as a whole.

Here is a conjecture concerning the line of reasoning that might be behind this "democratic" view. Influential individuals might perform such actions as *giving*

orders, or *bringing others to make their goals their own*, or any such acts. But they cannot simply *do* what requires the joint forces of many to be done, because this would require that the intentionality of the leader extend *directly* to the behavior of the subordinates. The other participants would be quite literally reduced to the leader's hands, feet, and other limbs: the behavior (body) being the subordinate's, the intentionality (mind) being the leader's. The subordinates would then, it seems, not be seen as acting on *their own* intention, in the execution of their *own* plans (however conformist they might be), since the intentional explanation of their behavior would point to the leader's wishes and intentions rather to their own. This is at odds with the conviction that all participants in plural actions, not just the leaders, have to be interpreted as agents. Therefore, the influence model cannot *literally* be true.

This worry, which seems to be quite widely shared in the relevant literature, closely resembles the commitment to individual intentional autonomy. But it is essential to be as careful as possible here. As I shall argue, the critical claim goes one step beyond the claim that each participating individual's behavior has to be interpreted as his or her own action. There is a further claim involved in this line of argument. To highlight both the proximity and difference to the assumption of individual intentional autonomy, I propose to call it the assumption of *individual motivational autarky*. I shall turn to the relation between the autonomy assumption and the autarky claim shortly. To introduce the idea of individual motivational autarky, let me just highlight the difference in focus. Whereas the principle of individual intentional autonomy states that individuals are (and should be interpreted as) *responsible* for – or *owners* of – their behavior, the autarky claim is a claim about the kind of *volitional resources* on which we might draw in ascribing agency to individuals. The term "autarky" usually refers to a closed economy, especially to the mercantilist ideal of an empire with no outside trade. The Greek word is composed of "autos", the self, and the verb "arkein", "to suffice", meaning self-sufficiency. This captures nicely what is at stake here on the level of *intentional* rather than economic resources. Individual motivational autarky amounts to the claim that in the last resort, *only the individual's own wishes, desires, projects, volitions, or whatever pro-attitudes he or she might have* are fit candidates to make sense of their behavior.

(C) Individual Motivational Autarky: Any interpretation of an individual's behavior has to bottom out in that individual's *own* pro-attitudes.

In other words, motivational autarky is the view that, on the basic level, individuals should be taken as acting exclusively on *their own* desires, plans, commitments, intentions and so on; loosely speaking, only in terms of the members of the acting individual's own "motivational set" is it possible to rationalize (or make good sense of) the individual's behavior.[10] This needs some further explanation.

[10] In Donald Davidson's words, "R is a primary reason why an agent performed the action A under the description d only if R consists of a pro attitude *of the agent* toward actions with a certain property, and a belief of the agent that A, under the description d, has that property" (Davidson 1963: 687, my emphasis).

Individual intentional autarky does, of course, not imply that *other individuals'* pro-attitudes can play *no role* in the interpretation of an individual's behavior. This would be obviously wrong, because people do not normally act *regardless* of what other people want. Quite often, people *do* take other people's pro-attitudes into account, and sometimes even *act on* other people's pro-attitudes. Thus it is clear that an interpretation of an individual's behavior should not methodologically treat that individual as *disregardful* of, or neutral towards, other people's wishes. Indeed this would be a great mistake, resembling some outdated economic models of human action. The idea of individual intentional autarky has nothing to do with such narrow-minded selfishness. Individual intentional autarky is perfectly compatible with the fact that other people's wishes can play an important role in how we act, and that we sometimes act *in accordance with* and even *on the base of* other people's wishes.

To put this differently, individual intentional autarky is not in conflict with the fact that action can be *other-regarding*. But – and this is the essential point – it imposes the following constraint: *if* individual A acts on individual B's pro-attitude, either of the following has to be the case: A has made B's wish *his or her own*, or A has some other appropriate pro-attitude, such as the wish to conform to B's wishes, or the wish to conform to the social norm of accommodating other people's wishes, or some such. For lack of a better term, let's call this the *other-directed pro-attitude condition*. It is always possible to do what the other wants, but if one does so, either of the following has to be the case: one has come to want it *oneself*, or one has some other appropriate pro-attitude such as the wish to conform to the other's desires, or the desire not to violate the appropriate set of rules of conduct, or some other kind of other-directed pro-attitude. In brief, motivational autarky is the claim that people cannot act on other people's wishes without having a volitional agenda of their own.

Just as with intentional autonomy, I introduced motivational autarky as a *methodological precept* rather than as an assumption about the ontology of action. I think, however, that there is a close link between methodology and ontology. To put it in a catch-phrase: methodology follows ontology. The question of whether or not we should stick to the rule of basing all intentional interpretations of an individual's behavior on a pro-attitude which we ascribe to that individual *herself* is ultimately settled by the question of whether or not there *are* such pro-attitudes at the base of the intentional infrastructure of the action in question. How is this question to be settled? Philosophers, as well as many non-philosophers, seem to take it for granted that the issue at stake is a *conceptual* one, and that motivational autarky is just as *essential* a feature of action as intentional autonomy (from which it is never clearly distinguished). This is to say that, according to the predominant view, an individual cannot be an *agent* without being motivationally autarkical. Motivational *heterarky* (i.e. the opposite of intentional autarky), just as intentional *heteronomy*, would displace that individual's agency. The reasoning behind this thesis seems to be the following: if the intentional interpretation of individual A's behavior were to bottom out in some of individual B's pro-attitudes rather than in any of A's own (other-directed) pro-attitudes, A's behavior would have to be interpreted *as B's action* rather than A's. Thus it seems that, insofar as A is an agent, she needs to be motivationally autarkical.

My claim is that this view is mistaken. I will not argue that there really *are* cases of motivational heterarky, even though I will present some evidence that this might actually be the case. Rather, my main aim is to show that while motivational autarky implies intentional autonomy, *the converse is not true*. Motivational autarky involves a *further claim*. Moreover, some forms of motivational heterarky are compatible with intentional autonomy. Let me first focus again on plural action. I have introduced the assumption of motivational autarky as the reason why most philosophers seem to think that the influence model of plural agency cannot literally be true. I certainly do not wish to deny that many (perhaps most) cases of action under the influence of another individual conform to the assumption of individual autarky. In these cases, the interpretation of these individuals' behavior has to bottom out in the respective individual's own (other-directed) pro-attitudes. But it is also true that there are some folk psychological views according to which no such additional other-directed pro-attitudes are needed in order to interpret an individual's behavior. This is particularly obvious in altogether unassuming cases of influence, especially in spontaneous, low-cost cooperative behavior. (We are now finally turning to the opposite extreme in the spectrum of the size of plural actions: from the Apollo program to smallest scale everyday cooperation). What is at stake here are simple patterns such as the following: holding a door open for a stranger, spontaneously helping a stranger to lift a baby carriage into the train, or moving aside a little on a park bench so that another person can find a seat, too (cf. Chapter 8 below). These are *social* actions, because they require cooperation, and they are *plural actions* exactly *insofar* as the helper's goal is the same as the individual's who is being helped (i.e. that the stranger pass the door, that the baby carriage be in the train, or that person P have a seat on the bench).

The decisive question is: do such cases conform to motivational autarky? I do not claim to have any conclusive evidence, but there are some reasons for doubt stemming from two sources: folk-psychology and the theory of empathy. From a pre-theoretic perspective at least, it does *not seem implausible* at all to assume that there need not be some wish to have another person sitting beside oneself, or a desire to conform to other people's wishes, or even just a particular disposition to conform to some set of rules, or *any such* pro-attitude, in order to move aside a little on the park bench (Paprzycka 2002). If I move aside, it might seem from a folk-psychological perspective that I do not do so because of anything *I* want, but I do so because of what *she*, the other, wants – and similarly for the other examples I have given. In this sense, the folk-psychological intentional interpretation of one's cooperative behavior does *not* bottom out in *one's own* volitions or pro-attitudes, but rather in *the other's*. Something similar seems to be true for certain kinds of acting under other people's influence, especially for some forms of obedience to authority, where people do not just give in to submissive desires of their own, but seem to have serious difficulty explaining to themselves why they conform to some other person's wishes.[11] Another line of argument that seems to suggest that there might

[11] Paprzycka (2002) mentions the case of Stanley Milgram's famous psychological experiments. I shall say more about this below.

be something wrong with the assumption of individual intentional autarky is the analysis of *empathy*. An important element of the *history* of the concept of empathy, from Theodor Lipps (1903) to current simulation theory, is the claim that there is a *direct connection* between the *understanding* of another individual's intentions, on the one hand, and *action tendencies* that are geared towards the same goal, on the other.

To sum up this argument, the effect of the dogma of motivational autarky is that it reduces our view of human interaction to cases in which there clearly *are* other-directed pro-attitudes. But folk psychology suggests that there are plural actions without such motivation. In these cases, folk psychology seems to allow for what I would like to label *motivational heterarky*. The intentional interpretation of such behavior does not bottom out in the acting individual's own pro-attitudes.[12] Remember that the "bottoming out" clause in the autarky claim allows for the fact that individuals often take other individuals' pro-attitudes into account, but requires that, in doing so, individuals act on a volitional agenda of their own. By contrast motivational heterarky is the claim that people may sometimes act on other people's wishes without having any volitional agenda of their own. I believe that there is much to say in favor of the assumption of motivational heterarky, and I hope that my remarks have been successful in raising some doubts concerning the universality of motivational autarky.

I will not defend and present any more evidence for motivational heterarky in this chapter. Instead, I set for myself a much more modest task for the remainder of this section, something I do hope to be *rationally* self-confident about. I will state and defend a claim concerning the *relation* between the assumption of intentional autonomy and the dogma of motivational autarky. My thesis is the following: our deep-seated conviction that each individual should be regarded as a responsible agent, and the widely shared assumption that the only intentional resource that can explain an individual's action are that individual's *own* pro-attitudes, are *two different claims*. In other words, it is possible to treat an individual as an *agent* without claiming that the interpretation of his or her behavior has to bottom out in his or her own pro-attitudes. Or, more precisely: while individual motivational autarky implies intentional autonomy, the converse is not true.

(D) Intentional autonomy does not imply motivational autarky.

I suspect that the main obstacle in the way of an adequate understanding of the role of intentional autarky in human action is that motivational autarky is mixed up with intentional autonomy. And, at first glance at least, it might indeed seem that the autonomy-claim and the autarky-claim amount to the same thing. After all, it does seem plausible that if an intentional interpretation of individual A's behavior were to bottom out in individual B's pro-attitudes rather than in any of individual A's own, the action in question would have to be attributed to B. A would be left no more than

[12] Views that are closely related to what I call motivational heterarky can be found in Roth (2006), Rovane (1998: Chap. IV), Paprzycka (1998, 2002). Needless to say, this paper owes a great deal to all of them. For further references, cf. Paprzycka 2002.

the role of a kind of a Manchurian Candidate, as it were, or an intentional zombie. But this view is mistaken (and the confusion of intentional autonomy and intentional autarky lies at the heart of our difficulty to understand the role of influence in plural action). It *is* possible to interpret A as acting on B's pro-attitudes without assuming other-directed pro-attitudes on A's part, and still interpret A's behavior as A's *own* action. Motivational heterarky does not *per se* compromise or displace an individual's agency.

The argument for my thesis is simple. It draws on the analogy between individual actions and those forms of plural agency which are at issue here, and on the distinction between motivation and intention. Take again the park bench case presented above as an example. Consider first a standard singular (or even solitary) version, in which A forms the intention to move aside on the park bench on the basis of some of his own desires. For the sake of the example, let's assume that a mild spring sun has come to shine on the side of the bench on which A is sitting. After a while, A is a little warm; pondering about whether to take off his jacket or to move out of the sunlight, he decides that he prefers the latter. So he moves to the shadowy part of the bench on the basis of his desire to cool down a little. It seems that this is perfectly sufficient to make sense of A's behavior. The intentional interpretation of A's behavior bottoms out in A's desire to cool down a little. A need not be ascribed any additional desire such as a desire to have his wish to cool down a little fulfilled. Still, A's moving aside undoubtedly is A's own action, even though A might not have a particular wish that he do what he wishes to do, but simply wants to cool down a little.

Let's now turn from the solitary to the plural action case. Assume for the sake of argument that it were in fact possible for A to form the intention to move aside on the bench on the basis of B's wish to sit down, without an additional underlying desire to conform to B's wishes. It seems hard to see why, in this case, the lack of some additional pro-attitude should now suddenly compromise A's agency, when it does not do so in the individual case. All that is needed to form an intention to move aside is some form of *awareness* of the other person's wish. It's still *his own action*, only now the intentional resources going into it extend beyond the range of A's own pro-attitudes. It's not that B somehow acts *directly* through A's behavior, bypassing and displacing A's agency. A's behavior does not have to be attributed to B's agency, rather than to A's; A does not behave on B's remote control. Rather, A's behavior still instantiates A's own action. A does not become B's *intentional zombie*, as it were, just because he acts on B's pro-attitude without there being any conforming pro-attitude from A's side involved. Behavior such as moving aside on park benches even when one does not have any particular wish to have another person sitting beside oneself, or to conform to other people's wishes, or even to conform to the norms of propriety, is not a form of intentional zombieism – not even a mild one. Rather, it is a matter of simple *politeness* (even though not all heterarkical behavior is of the nice, beneficial kind, as we shall see shortly). Heterarkical agents *do intend what they do* (e.g., move aside on the bench), but the chain of intentional interpretation leads beyond what's in their own solitary motivational set. These agents are intentionally autonomous, but not motivationally autarkical.

If this is true, if individual intentional autonomy is conceptually independent from the dogma of motivational autarky, and if there is such a fundamental difference between the two, the question arises: how come they have always been lumped together? *Why do we tend to mix up the idea of being the agents responsible for our own behavior with the apparently very different idea that in the last resort, only* our own desires *are fit candidates to make sense of our behavior?* In short, my answer is this: it is because, in *our culture* at least, motivational autarky describes the way people are *supposed to be* (and see themselves). Being the one and only ultimate source of the intentional infrastructure of one's own behavior may not be a conceptual feature of agency, but in our particular culture at least, it is a very basic and extremely strong *normative ideal*. While a person's explaining her actions in terms of another person's intentions is quite frequent in everyday talk, we tend to press for "deeper" explanations, and even to react *embarrassed*, if a person fails to come up with some pro-attitudes of her own in explanation of her behavior. It is as if such a person had failed to conform to our idea of selfhood, and it is very tempting to blame this on her way of describing her action, rather than on the structure of her action. People, we seem to think, really *shouldn't be doing things just because other people wanted them to be done*, without thereby conforming to any of their own wishes – and insofar as we regard them to be fully developed selves, they just *can't*. Therefore, we tend to believe that in such cases, there has to be something wrong with their interpretation of their action.

This negative evaluation of heterarkical behavior might surprise, especially since most of the presumptive cases of motivational heterarky discussed above are of a rather beneficial, pro-social kind (think of the park bench example). But even in such cases, heterarky it is not well regarded. People are welcome to assist other people, but in our culture at least, it is believed that they should be performing such acts *because they wanted to be of help*, and not just because other persons wanted those acts to be done. Moreover, there are distinctly negative examples of motivational heterarky. A vivid illustration is provided by Stanley Milgram's (1974) famous psychological experiments (here, I follow a hint given by Paprzycka [2002]). Remember that Milgram's test subjects – perfectly decent ordinary people – proved willing to administer deadly electroshocks to innocent others, just because they were told to do so by some authority figure. There were neither financial incentives nor sadistic inclinations involved. So how come those people did what they did? Of course, it is always possible to assume that people acted on some *desire* to conform to the authority figure's wishes, or some desire to be a "good" and obedient collaborator, or some such pro-attitude. Based on repeated interviews with his subjects, however, Milgram himself gives another explanation for his stunning results. He explains his test subjects' behavior with what he calls an "agentic state".

An agentic state, Milgram explains, is a condition in which a person sees herself as *acting on another person's desires rather than on any of his or her own* (Milgram 1974). The most convincing evidence for the existence of agentic states is the fact that in the interviews carried out immediately after the experiments, and again some months after the event, many test subjects proved to be utterly unable to *explain to themselves* why they acted in the way they did, and did not come up

with any compliant inclinations in explanation of their behavior. The reason might be that there really *were* no such compliant pro-attitudes. It strikes the reader of the subjects' statements printed in Milgram's book that this utter cluelessness concerning the deeper motivation for their action is even true of some of those test subjects who explicitly accepted full *responsibility* of what they proved capable of doing during the experiment, and were not just looking for excuses. It might well be that this cluelessness stems from the fact that the subjects were looking for the reason for their action in the wrong place: in their own "motivational set", instead of in the authority figure's.

In his book, Milgram tends to dismiss agentic states as some sort of illusion; moreover, he depicts agentic states as an *unusual* condition that requires the presence of authority. As is well understandable from his experiments, he sees agentic states as morally utterly condemnable. Thus the normative ideal of intentional autarky becomes very clear in Milgram's depiction of the fatal consequences of agentic states. By contrast to Milgram, I propose to consider three things: first, agentic states might not be simply a matter of self-deception; second, motivational heterarky might be a *normal* condition rather than an exception, which, third, may lead to morally disastrous consequences under conditions such as those examined by Milgram, but can also be very beneficial under such circumstances as those found in public parks, airports, and railway stations (think of the beneficial and cooperative examples of motivational heterarky mentioned above). In short, I will not pass any judgment on whether or not we should hold on to our ideal of motivational autarky. What is certain, however, is that we cannot even start to discuss the question of whether or not motivational autarky is indeed an ideal worthy of defense, if we continue mixing it up with intentional autonomy. Because intentional autonomy is a constituent of *any* action, it is not to be changed. By contrast, motivational autarky is a cultural ideal, which we may or may not want to uphold.

With this result, let us finally come back to the question of the role of influence in plural agency. If we accept the possibility of motivational heterarky, it seems that the problem diagnosed above simply vanishes. We do not have to deny the possibility that sense can be made of plural actions *as a whole* in terms of the intentionality of the leading individual, as is so often done in everyday parlance. If we discontinue mixing up intentional autonomy with motivational autarky, it becomes obvious that to base an intentional interpretation of the participating individual's behavior in question on the leading figure's volition does not mean to bypass or to compromise the other participating individual's agency. We may ascribe the leader an *intentional authority* over what is going on by way of an intentional interpretation of the behavior in question that bottoms out in the leading figure's volitions, without thereby divesting the other participants of their own individual agency. The other contributors can still be seen as *agents*, with their behavior instantiating *their own action*. The leading figure's claim to the entire action does not necessarily disregard the other participants' own individual agency, because it does not interfere with the other participants' ownership of their individual contribution to the plural action. That it really was *Caesar* who defeated the Helvetii in the battle of Bibracte (in the sense that an intentional interpretation of the movements of the roman legions

bottoms out in no other than his attitudes), does not contradict the fact that even the most obedient and servile soldier of his beloved tenth legion threw his javelin *himself*. Thus there is ample room to take the commonsensical notion of the role of influence, power, and volitional openness more seriously in the theory of plural agency, without letting go of the idea of the fundamental intentional autonomy of *all* participants, not just the leaders. Here, as in the above case of the collective agent model of plural agency, action theory should be more accommodating towards common sense.

§4 Intentional Individualism

In many cases, some collective agent provides the solution to the plural agent problem. In others, some individual's influence and authority does the job. While both models should be taken more seriously in action theory, it seems clear that the extension of either of these models is limited, that there are cases of plural actions that can be fitted into neither of them, and that both types of plural agents presuppose plural action of another kind.[13] Consider again the following example. If you and I go for a walk together, this is clearly a case of a plural action – the action requires more than one participant, and the two of us will pursue the same (token) goal, i.e. walking together. I have argued that this case cannot be fitted into the collective agent view, because collective agency involves a kind of agency that is *different* from that of the participating individuals, which does not seem to be the case here: if you and I go for a walk together, there are *two* subjects involved in the case, not one (there is no collective agent walking all by himself), and not three (there is no additional collective subject escorting the two of us through our walk). Thus the collective agent view does not cover this case. It also seems to be futile to try to fit it into the influence model. Each of us will be walking with the other, but none of us has a claim to ownership with regard to our walking. Our walking is something *we own together:* in such cases, *ownership is shared*. This brings us, finally, to the third and last common-sense concept of plural agency. In this last view, the plural agent is not *one* agent – neither *one individual*, as in the influence view, nor *one collective*, as the collective agent view has it. Rather, the plural agent(s) are *many*: acting jointly, as it were, or hand in hand, in pursuit of the one shared goal.

I believe that this model of plural agency – I shall call it the *teamwork view* – is *the most basic one*. Teamwork is *presupposed* in the collective agent view, insofar it is only by virtue of teamwork that there are any collective agents at all; for there to

[13] I have argued above that, for collective agents to emerge from the agency of individuals, there has to be some communication going on between these individuals, which is plural action. A similar point can be made with regard to power and influence. In most cases, power is based in – or generated by – some form of *collective acceptance.* This is a shared intentional attitude, which is typically expressed in some form of declaration, or affirmed in some other form of public expression. This in turn is a plural action of the third type, i.e. teamwork.

be a collective agent, individuals have to *act jointly* in pursuit of the goal to *create* and *maintain* a collective agent. Also, it seems that most cases of the influence type of plural agency can also be modeled on the teamwork view. If it seems correct to ascribe the Helvetii's defeat in the battle of Bibracte to Cesar, it is no less correct to ascribe this action to *the Romans*, or to those Romans active in the course of the events, acting jointly as a team under Cesar's leadership. Thus it seems that the teamwork view is much more than just one view of plural agency among others. It is the bedrock of plural agency, and should therefore be the main focus of any theory of plural action.

But this model, too, has its difficulties. The most obvious problem any theory of teamwork will have to cope with is that of reconciling the *unity of the action* on the one hand with the *plurality of agents* on the other. In the current debate, this is *precisely* what the *concept of shared or collective intentionality* is meant to do. The claim is this: many people can intend *one and the same action* precisely insofar as they *share* the respective intention. The problem, however, is that it is somewhat unclear what it means to share an intention. Cakes and cars can be shared – one (token) cake, many pieces, one (token) car, many users – but intentions? What can talk of "sharing" possibly mean in this context?

Looking at the debate on collective intentionality that has evolved over the last 20 years, it becomes obvious that most authors tend to understand the sharing of intentionality not in the straightforward sense, but rather as a metaphor. According to authors such as Raimo Tuomela (as read by John Searle), John Searle himself, and Michael Bratman, there is no single (token) shared intentional state that is behind the joint intentional activity, but many intentional states instead, individual intentional states that are marked out from those involved in the case of solitary or singular agency in that they are either of a special *form* (Searle), *mode* (Tuomela), or *content* (Bratman), providing the "glue" for collective intentionality. In other words, the existing accounts of collective intentionality tend to be of a *distributive* kind. I call *distributive* those conceptions of collective intentionality which claim that, whenever people share an intention, each individual has his or her own intention, and that there is no such thing as one (token) intentional state that is shared by the participants in the straightforward sense of the term.[14] In other words, distributive conceptions of collective intentionality are marked by what I propose to call *intentional individualism*.

(E) Intentional Individualism: Any interpretation of an individual's behavior has to be given in terms of *individual* intentional states.

[14] In the second chapter of his *Analysis and Metaphysics* (1992), Peter F. Strawson introduces a distinction which is important to correct John Searle's influential misunderstanding of Tuomela's position. Strawson distinguishes between *reductivist* and *connectivist* analysis. In contrast to analyses of the reductivist kind, connectivist analyses do not identify independently existing "building blocks", or "atoms", but rather elements that might, for their existence, be dependent on each other. Insofar as this is true for Tuomela's analysis of shared intention, his position is not distributive, contrary to what Searle's reading suggests. In the meantime, Tuomela has repeatedly endorsed a non-reductivist reading of his position (e.g. in Tuomela 2007).

I have put intentional individualism like this so that it is a much weaker claim than individual motivational autarky. It constrains the class of possible mental states required to make sense of an individual's behavior to some *individual subject's pro-attitudes* (which could be *either* the respective individual's own – this is the case of individual motivational autarky – or any other individual's). Needless to say, most philosophers of collective intentionality implicitly accept not only intentional individualism, but individual motivational autarky, too, claiming that the pro-attitudes in terms of which sense should be made of the behavior of any individual participating in a plural action should not only be *individual* pro-attitudes, but that respective individual's *own* pro-attitudes. Having discussed some of the problems of the dogma of motivational autarky at some length in the last section, however, let's focus on intentional individualism here. It might seem that intentional individualism is so obviously true that it is not in need of further substantiation. It might appear that to assume that there is collective ownership of an action that cannot be ascribed to a separate collective subject, but that is *shared* among the participants intending the action *jointly* in the straightforward sense, would seem to amount to some implausible *fusion of mind*. Most philosophers of collective intentionality think that the idea of a non-individual mind is so terribly and obviously mistaken, that there is no need for further argument.[15]

Before examining the hidden background of this almost universal endorsement of intentional individualism, let me first say a word about why I think it might be problematic. As I have said, intentional individualism forces us to adopt a distributive conception of collective intentionality. The problem with distributive conceptions of collective intentionality – at least with those that have been put forward so far – is that they tend to be *circular*. The objection of circularity points out that whatever *individuals* intend when they share an intention, already *presupposes* the shared intention.[16] In other words, the dogma of intentional individualism makes it impossible to understand the element of *intentional commonality* that seems to be *presupposed* whenever people form an intention to participate in joint intentional activities. The circularity issues of the existing distributive conceptions of collective intentionality lends some plausibility to the conjecture that intentional commonality is indeed *irreducible*, and cannot therefore be captured by a distributive conception of collective intentionality. Intentional commonality, as I propose to use the term, is incompatible with intentional individualism; it implies sharing an intention in the straightforward sense of the word: one (token) intentional state, many participants.

---

[15] Thus Searle – without providing any further argument – shuns such "perfectly dreadful metaphysical excrescences" (Searle 1998b: 150). Tuomela, in turn, dismisses non-individualistic conceptions of the mind as "spooky" (Tuomela 1995: ix, 5, 353, 367). For a closer examination of the role of the group mind in collective intentionality analysis cf. below Chapter 2.

[16] ***I*** can intend to do my part in a plural action x *as my part* only if *we* intend to do x; thus my intending to do my part as my part is no independent "building block" of collective intentionality, as a superficial reading of Tuomela/Miller's (1988) account has the authors claiming, but rather an element of a holistic intentional structure. Similar points can be made with regard to Michael Bratman's and John Searle's distributive accounts of collective intentionality (for a detailed analysis cf. Schmid 2005c).

If this is true, if there really are some more or less obvious problems understanding the structure of collective intentionality along the distributive line, and if this is due to intentional individualism, the question arises: why do most philosophers of collective intentionality simply take intentional individualism for granted, and not even think that it is necessary to provide an argument for its validity? My conjecture is the following: just as motivational autarky is usually mixed up with intentional autonomy, the possibility that intentional individualism might not be implied by individual intentional autonomy is overlooked. I believe that the worry in the back of the mind of the distributive philosophers of collective intentionality is that, if intentional individualism were not true, the individual participants would be deprived of the ownership of their contributive action, thus comprising their agency. This worry, however, is unsubstantiated.

(F) Individual intentional autonomy does not imply intentional individualism.

It is possible to interpret several individuals as sharing one intentional state (in the straightforward sense of the term), and still interpret these agents as the owners of their contributive action. The argument follows precisely the same line as the one we used to establish the independence of intentional autonomy from motivational autarky in the last section. The upshot is this. Even within an intentional interpretation of a given behavior that appeals to a *shared* intentional state rather than to any of the participant's own *individual* intentional state, we may still interpret the participating individuals as intending their contribution *individually*, and as owners of their individual contributive actions. The fact that their individual contributive intention is *derived* from a shared volition does not undermine the individuals' ownership of their contribution.

Intentional commonality does not compromise the participating individuals' agency. It is not the case that some group mind *displaces* the participating individuals' agency if those individuals were to act on a shared intention. Each individual still intends to do his or her own part *individually*, and is thus the owner of his or her contribution, but this participatory individual intentionality is derived from an intention that is not individual, i.e. from the *shared* intention to carry out the plural action in question together. Thus the chain of intentional interpretation of the individual behavior in question leads beyond what is intended individually – without thereby flying in the face of the fundamental idea that each individual participating in the process is an agent in his or her own right.

I think that once it becomes clear that we might drop intentional individualism without letting go of intentional autonomy, the urge to exorcise the group mind from collective intentionality analysis vanishes. We can be more relaxed with regard to the non-individualistic conceptions of the mind that are so pervasive in much of the earlier history of thought, and from which we might still learn a lesson or two about what it means to act jointly, as a team.

§5 Plural Agency and Methodological Individualism

Let me now wrap up the line of argument developed in this chapter. Where does the resistance against conceptions of plural agency that is so pervasive in action theory come from? I suggested that the attitude underlying much of this resistance is methodological individualism. The term "methodological individualism" was coined 100 years ago (Schumpeter 1908: 88–98). Schumpeter introduced the term to label the views he shared with Max Weber. Ever since Schumpeter coined the term, the issue has kept coming up in social philosophy, usually in cycles of about 20 years or so (cf. Udehn 2001). In each round, the controversy had a somewhat different focus. Under the title 'methodological individualism', issues as different as the limits of social planning, the relation between social action and social structure, and the role of collectivity concepts in social explanation have been discussed. I believe that the right way to celebrate the centenary of methodological individualism would be finally to come back to the heart of the matter.

At its (historical) heart, methodological individualism is about *plural action*, and more precisely, it is the claim that there are no plural agents. At the basic level of intentional interpretation, all agency has to be treated as singular agency: this is how one might summarize the central precept of methodological individualism. In other words: plural actions, as we encounter them in social life, should be ascribed to singular agents. This does, of course, not mean that there is no *social action*, i.e. that all action is (in the terminology developed above) of the solitary kind. Methodological individualists are well aware that many actions presuppose for their possibility the actions of other individuals. And it does not mean that there are no plural actions. Methodological individualists do not have to deny that sometimes, individuals act jointly in pursuit of the same goal. All that is claimed by methodological individualists is that these actions do not require any particular notion of agency, i.e. that it is enough to assume singular agents for the purpose of the interpretation of plural actions.

I have used the principle of rational intentional self-confidence to cast some doubt on this view above. If a plural action can indeed be legitimately interpreted *as one action* (and not just as an aggregate of many actions), singular agency simply will not do: we need a notion of plural agency (I labeled this the "plural agency problem"). Common sense has no difficulty providing suitable candidates for this role. The problem, however, is that these common sense interpretations of plural actions are not well received in action theory, and I suspect that a wide-spread, more or less tacit commitment to methodological individualism is to blame.

In this chapter, I argued that methodological individualism (qua singularism about plural agency) rests on three mistaken conclusions drawn from one valid insight. The basic insight is *individual intentional autonomy*: however plural an action might be, each participating individual's behavior has to be interpreted as *his or her own action*. We are, in other words, not intentional zombies. The three mistakes are the following: first, contrary to what methodological individualists seemed to think, individual intentional autonomy does not rule out the irreducibility of collective agency. Second, methodological individualists are mistaken in assuming that

individual intentional autonomy rules out what I called *motivational heterarky*, i.e. behavior whose intentional interpretation bottoms out in pro-attitudes that are not the respective individual's own. Some plural actions can be ascribed to influential individuals or authority figures, without thereby bypassing the other participating individuals' ownership of their contributive actions. And third, methodological individualism is wrong if it amounts to endorsing the view that individual intentional autonomy is in conflict with robust notions of *intentional commonality*. We need a solid conception of intentional commonality to solve the circularity problems encountered by the existing distributive notions of collective intentionality. Such a conception can be compatible with individual intentional autonomy. This insight should help to overcome the widespread fear of non-individualistic conceptions of the mind, and lead to more adequate theories of *teamwork*.

While there might be some *cultural* and *historical reasons* for the fact that the assumptions of autonomy, autarky and individualism usually come as a package, there is *no reason* why we shouldn't *unpack* methodological individualism and start to think about which items to keep and which to throw out. Well understood, this is *not to say* that those ingredients of methodological individualism which we might find unfit for the purposes of the theory of plural action might not turn out to be useful for some other purpose (e.g. as cultural ideals). As we know, the extent to which human coordination and cooperation is achieved by plural action varies from group to group, from society to society, and from time to time. While I am quite convinced that individual intentional autonomy is *universal*, I think that the same is not true for autarky and individualism. It might turn out that there is much more autarky and much more individualism in *some* societies than in *others*.

This brings me to my final point. It is well known that in the paper in which the term "methodological individualism" was first introduced into the English language, Joseph Alois Schumpeter himself limited its validity. It is often quoted – by Kenneth Arrow (1994), among others – that Schumpeter (1909) says that the social can sometimes be considered "as if" it were an "independent agency". Nobody seems to have noted so far, however, that there is yet another, much blunter limitation stated in this paper. Schumpeter goes as far as to admit that methodological individualism is a relative principle which should not be applied to a certain type of society: it should not be applied to *communism*. Writing in 1908 (the paper was published in 1909), there was no way for Schumpeter to know what was to come under this label. So what did he mean with the term communism? He meant this: a society in which there are not just individual wants, but shared wants, too, and where there is joint action based on these "social wants". Let me quote a passage from Schumpeter's paper:

> The only wants which for the purpose of economic theory should be called strictly social are *those which are consciously asserted by the whole community*. The means of satisfying such wants are valued not by individuals who merely interact, but by all individuals *acting as a community consciously and jointly*. (Schumpeter 1909: 216)

This means plural agency, and indeed it means plural agency of the fundamental *teamwork* kind. On the one hand, it might be true that there was much less teamwork in later socialist societies than Schumpeter could ever imagine. Yet on the

other hand, there are, without doubt, many more genuine teams at work in capitalist societies than individualists like to think. This is not just the case in large-scale ventures such as the Apollo-program, but, above all, in altogether unassuming everyday interactions. Plural Action is an important part of life. And if Schumpeter is right, it cannot be adequately understood within the framework of methodological individualism.

Chapter 2
Overcoming the 'Cartesian Brainwash'

Beyond Intentional Individualism

Among the many reasons why John Searle is important in the debate revolving around the structure of collective intentionality is the fact that even though the history of the analysis of collective intentionality has roots that go further back,[1] it was him who coined the term (Searle 1990). The following chapter sheds some critical light on a feature which Searle's account shares with most of the received accounts of collective intentionality. It is argued that fear of the group mind has played a fateful role in the early stages of the current debate by driving most philosophers of collective intentionality into accepting one or another version of intentional individualism.

§6 Collective Intentionality Without Collectivity?

On his way towards a "*general* theory" (Searle 1998a: 161), John R. Searle has recently started to venture into what he likes to see as a new field: 'Philosophy of Society'. In some of the papers surrounding his *Construction of Social Reality*, Searle envisages this discipline to be centered on how the individual and society relate to each other (1997b: 103, 1998b: 143). Of course, this is hardly a new question. It has been the topic of many a philosophical debate and controversy at least since it became common practice to refer to single human beings as individuals. In the second half of the last century, individualism has become the dominant view of the basic structure of this relation. In much of social science, it has come to be widely held that explanations of social phenomena have ultimately to be given in terms of individual actions (cf. Popper 1962: 98). Let us label this the orthodox view. In its reductivistic form, the social is *nothing but* an aggregate of individuals who decide over the alternatives they believe to be available to them in the light of whatever preferences they have; in the last resort, it is each individual deciding for him- or herself. Thus the social is *secondary* as compared to the intentionality of the single

[1] The direct roots of the concept are in Robin George Collingwood's *New Leviathan* (1942), where Wilfrid Sellars picked up the term, which was then analyzed by Raimo Tuomela (1984; see also Tuomela/Miller [1988]) who initiated the current debate.

H.B. Schmid, *Plural Action: Essays in Philosophy and Social Science*, Contributions to Phenomenology 58, © Springer Science+Business Media B.V. 2009

individuals. It appears that we do not have to presuppose collectivity concepts such as 'group' or 'community' in order to analyze what it means for an individual to optimize his or her expected utility. Collectivity concepts enter orthodox explanations of social phenomena only insofar as they are either the direct *object of individual intentions*[2] or among the *unintended consequences of individual actions*.[3] In both cases, individual intentions and actions – and *not* collectives – are what social science is about. For in this view, the social does not reach down to the form and structure of intentionality and action itself.

In spite of the near total "Triumph of the Individual" in social science, some opponents and contesters survive. In philosophy, it is widely held that interpretation (and thus interaction) is a precondition for there to be mental states with intentional content. Social externalism even seems to have become the mainstream position. In social and sociological theory, too, some 'heterodox' strands persist in opposing the allegedly 'atomistic' picture of human agency and intentionality. Heterodox philosophers of society emphasize that in many respects, individual intentionality is more deeply imbued with sociality than the 'orthodox' view has it.[4]

At first glance, it might appear that the heterodox view receives further support from one of the most exiting recent developments in analytic philosophy of intentionality and action. The works of, among others, Raimo Tuomela, Margaret Gilbert, and Michael Bratman, together with John Searle's own contributions, have substantially broadened our understanding of intentionality and action. After the traditional concentration on the individual intentionality of single agents, the focus of attention has now shifted to an analysis of what it means to intend and act *together*, a phenomenon which, by and large, had received only marginal notice in the earlier philosophy of intentionality and action.[5] By now, it is well-established and widely recognized as a fact that intentionality is not exclusively a matter of the personal beliefs, desires and expectations of individuals. What makes our intentionality and our actions *social* is not just that from time to time, we make each other the object of our individual intentions or expectations. Rather, intentionality is *in itself* something human beings can *share*.

This, it seems, adds a new facet to the question not only concerning the relation between the individual and the collective, but more generally concerning the relation between the mental and the social. In his *Common Mind* ([1992] 1996), Philip Pettit has distinguished two dimensions of that relation: the "vertical" issue concerning the question of whether or not collective forces trump intentional regularities, and the "horizontal" issue of whether or not – or how far – the capacities

---

[2] For a classical expression of this view see Weber ([1921] 1980: 7). For Weber, collectivities can be 'real' only in the sense that they are *believed to be real* by the individuals.

[3] For this view see Elster (1989).

[4] The most commonly known 'heterodox' strand in social ontology is Communitarianism; cf. Sandel (1982).

[5] Early exceptions to the rule can be found in Phenomenological Philosophy (see, e.g., Walther (1923); for a more "holistic" view see Stein (1922: 116–267); Heidegger ([1928/29] 1996: 83ff.); Sartre ([1943] 1991: 464ff.)).

that mark us out as human beings essentially depend, for their existence, on social relation. Pettit proposes to label the first issue the controversy between individualism and collectivism, and to use the Greek counterparts of these terms – atomism and holism – for the contending positions in the second issue. Famously, Pettit argues for individualism (and against collectivism) in the vertical dimension, and for holism (and against atomism) in the horizontal dimension. But he does not, in this book, address the particular way from which Tuomela (1984, 1995), Searle (1995), Gilbert (1989) and Bratman (1999) had started to approach the question of the social just a few years before Pettit's *Common Mind* first appeared in print. Comparing the way in which Pettit treats the question of the "commonality" of the mental with the way in which these other philosophers approach the topic makes the novelty of the analysis of collective intentionality all the more apparent. Pettit addresses the traditional issues that have been on the agenda of research in social ontology at least since social externalism first appeared on the scene; with Searle et al., a new perspective has opened up – a perspective that largely seems to be independent of the earlier debates. In order to say that intentionality can be genuinely *shared* one need not assume that metaphysical socialism about intentionality is true, i.e. that *any kind of intentionality* is a social fact, let alone that collective forces compromise the intentional psychology of individuals as some extreme collectivists had it.[6] Thus collective intentionality does indeed open up a new perspective on the relation between the individual and the collective levels.

Upon a closer look, however, it appears that heterodox philosophers of society should not put their hopes for support from the theory of collective intentionality too high. The main protagonists of this movement do not seem to think that their novel approach to the structure of intentionality and action should open a new perspective on the basic structure of the relation between the social and the individual. By and large, the orthodox account is left intact. Raimo Tuomela virtually treats groups as ontological non-entities because in his view, "groupness" is, as he puts it, "in the last analysis attributed to individuals" (Tuomela 1995: 199). It seems that in his account, the basic structure of we-intentionality does not *per se* presuppose collective entities such as 'groups' or 'communities'. For an individual to we-intend it is, following Tuomela, not necessary that other agents actually exist, much less that there is an actual we-group.[7] Margaret Gilbert, meanwhile, has repeatedly claimed to go "beyond individualism" (Gilbert 2000: 3). Yet in her book *On Social Facts*, she explicitly bases her analysis on a concept of the individual that "does not require for its analysis a concept of a collectivity" (Gilbert 1989: 435ff.). The conceptual basis of her account of "joint commitment" consists of nothing but conditional personal commitments (Gilbert 2002). Michael Bratman, in turn, calls his own theory "reductive in spirit" because he takes shared intentionality to be analyzable "in terms of attitudes and actions of the individuals involved" (Bratman 1999: 108). Last but

[6] In the meantime, Pettit has taken his stance on that new topic (Pettit 2005). It seems, however, that Pettit does not recognize that the construction of collective agents presuppose collective intentional activity from the side of the participants.

[7] Tuomela (1991: 254). See also Hindriks (2002).

not least, Searle himself not only hastens to declare that his account of collective intentionality is fully consistent with methodological individualism (Searle 1990: 406). He also stresses the ontological primacy of what goes on in the individual mind over the existence of the group by pointing out that, "ontologically speaking, collective intentionality gives rise to the collective, and not the other way around" (Searle 1997a: 449).

Throughout this debate, actual collectivity is, it seems, held to be methodologically and ontologically secondary to (and derivative from) the mental activity of the single individuals involved in collective intending. The actual existence of a we-group is seen as a more or less contingent by-product of the intentionality of individuals. From a heterodox point of view, this debate gives the impression of an attempt to account for the structure of collective intentionality without letting any genuine collectivity enter the scene. "We-ness" is the topic, yet at the same time it is stressed that it is a feature of individuals – and not of an actual 'we'. Annette Baier expresses the dissatisfaction heterodox philosophers of society might feel rather drastically when she takes the current debate on collective intentionality to prove that Descartes has thoroughly brainwashed us (Baier 1997b: 18).

§7 The Specter of the Group Mind

Before taking a closer look at this sweeping diagnosis, I would like to highlight a rather somber figure that is haunting this debate, and that seems to have played a crucial role in the formation of its individualistic setting. It is the specter of the *collective subject*, or *group mind*. Its importance in this debate seems to stem from a rather innocent-looking assumption. Where there is intentionality, it is said, there has to be somebody who 'has' it – the good old subject. Now if it is claimed that there is such a thing as *collective* intentionality, and that collective intentionality is to be distinguished from *individual* intentionality, the conclusion seems to force itself on us that it has to be, not *single individuals*, but *collectives* themselves that 'have it'. And for collectives to have intentions, some sort of a 'collective mind', some 'group mind', seems to be required, something hovering over and above the minds of the individuals involved. To the untrained eye, at least, this apparent implication of the very concept of collective intentionality does not look very appealing. Thus, among the protagonists of collective intentionality, it was originally widely agreed upon that there is no mind over and above the minds of individuals (ironically, it was Pettit who eventually showed that one need not be all that worried about ascribing a "mind of their own" to groups [2003]). While the question of whether or not (and, if so, in what sense) collectives can be agents in their own right remained to some degree controversial even in the early analyses, it seemed almost universally agreed upon that it is unacceptable to treat collectives as 'subjects' of intentions and actions in the *ordinary* sense in which individuals are the bearers of

their intentionality.[8] Even where the notion of the collective subject was stripped of its mentalistic content, it still did not quite appeal to most philosophers of mind and action, because it seemed to be associated with collectivistic[9] (or even totalitarian[10]) notions of the social. If it is to the *collective* rather than the individuals that intentionality is ascribed in order to make sense of an observed behavior, the participating individuals seem to be no more than organs, i.e. mere instruments, and this seems to contradict our idea of individual intentional autonomy.

Thus it seems quite understandable that the above-mentioned philosophers of collective intentionality set themselves the task of showing that collective intentionality is possible without there being a group mind (let alone some collective consciousness) involved. The specter of the group mind (or collective subject) had to be exorcised, and one can identify two different ways in which this was done. The softer way – it might look more like psychotherapy than like hard-core exorcism – was chosen by Margaret Gilbert, Raimo Tuomela, and, perhaps, Robert Sugden. In these analyses, some sort of collective subject is admitted to the theory, but it is domesticated so as to be consistent with an otherwise thoroughly individualistic conceptual framework. Here, either some rather strong sense of membership to a collective (Sugden[11]) or some softened and modernized version of the collective subject itself (Gilbert,[12] Tuomela[13]) is made part of the theory. At the same time, however, the collective subject is solidly founded in the intentional autonomy of individuals by reducing the collective subject either to sets of individual intentions[14] or to the reflective self-understanding or self-categorization of the single participating individuals qua members of the team.[15] The tougher way of dealing with the specter of the group mind was simply to treat it as an abominable collectivist idea that has to be banished from the theory of collective intentionality straight away. On this tough line, the group mind is exorcised either by stating that all intentionality involved in collective intending is exclusively the intentionality in the minds of the participating individuals, or by making the somewhat different point that the intentionality individuals "have" when participating in collective intentionality is basically a form of their personal intentionality. These are the strategies that were chosen by Searle and Bratman, respectively.

---

[8] Edmund Husserl's theory of "higher order-persons" gives an illustrative example of the difficulties that any attempt to apply the model of the individual subject to collectives will face (cf. Schmid 2000: 17–27).

[9] Cf. Emile Durkheim's concept of the "collective consciousness" ([1898] 1994).

[10] Cf. Hartshorne (1942).

[11] Sugden favors a concept of membership "in something like the old sense in which arms and legs are members of the body" (1993: 86). This reminds of the Aristotelian view of the relation between society and the single human beings (see Aristoteles: *Politics* 1253a), a view that – at least at first glance – appears to be inconsistent with our modern view of the single human beings as *individuals*.

[12] Cf. Gilbert's concept of the "plural subject" (1989).

[13] Cf. Tuomela 1995: 231).

[14] See Tuomela (1995).

[15] See Gilbert's "Simmelian" account in Gilbert (1989: Chap. 4). In Sugden's view, "a team exists to the extent that its members take themselves to be members of it"; Sugden (2000: 192).

Before turning to this, let me state a general observation. Looking from afar at how the group mind was dealt with in the early stages of the debate on collective intentionality, it might appear that the theory of collective intentionality was caught in a dilemma, or rather, stuck in some kind of double-bind, as it were. On the one hand, the aim was clearly to break with individualism in the sense of the orthodox limitation to purely individual intentionality, which is recognized as being overly restrictive and unfit for our understanding of the social world. On the other hand, however, individualism (in the broad sense of an emphasis on the role of the individual) seemed to be the only effective defense against the specter of the group mind. Thus, in a sense, the theory of collective intentionality had to reject and to endorse individualism at the same time. How was this apparent tension dealt with? In exploring this question further, and in showing how detrimental this constellation was for the further development of the analysis of collective intentionality, I shall concentrate on the hard-line accounts against the group mind, i.e. Searle's and Bratman's.

Following up on Baier's statement concerning the 'Cartesian brainwash', a short remark on Descartes is in order. What is 'Cartesian' about how such authors as Searle and Bratman put their respective analyses of collective intentionality? Let us start with the most obvious sense in which there is something non-social about Descartes' venture. In his *Meditationes*, Descartes makes clear that his aim is to contemplate his own mind in "lonely withdrawal"[16] from society. Thus it is hardly surprising that he comes up with a rather under-socialized account of the mind. There are two ways in which his view of the mind is individualistic. Firstly, the mental comes exclusively in the form *ego cogito* – and not, as Charles Horton Cooley would have already liked to have it, in the form *nos cogitamus*.[17] This is to say that Descartes' account is individualistic in that it restricts intentionality to the form "I intend", "I think". It simply does not seem to have crossed Descartes' mind that there could be intentionality in the first person *plural* form, too. I shall refer to this first version of individualism with the term *formal individualism*, for what is at stake here is the *form* of intentionality.[18]

In a second and quite different sense of the term, Descartes' account is individualistic in that he portrays the individual mind as a *solitary place of representations*. Whatever the contemplating self finds in its mind is, following the view that was first articulated by Descartes, *structurally independent* of any relation to anything outside that individual mind. There is no telling whether a belief does or does not represent a real state of affairs just by reflecting on that belief *qua* mental state. Even the existence of some *genius malignus* who has the power of making me be mistaken in my beliefs could not thereby bring about the slightest structural change in my intentionality. "Being in a state with specific cognitive content does not essentially involve standing in any real relation to anything external" (Segal 2000: 11). In

[16] Cf. Descartes, René: *Meditationes de prima philosophia*, First Meditation, §3.

[17] Cf. Cooley ([1902/05] 1956: 6).

[18] There are other terms in use for this kind of individualism. Kay Mathiesen (2002) proposes the term "phenomenological individualism" as opposed to ontological individualism.

the current debate, this view usually goes under the label 'internalism', but since internalism is usually taken to include a non-Cartesian account of the relation between the features of our physical brain and our mind, I shall use the term *subjective individualism* instead.[19] This is the second sense in which Descartes' view of the mind is individualistic. As opposed to formal individualism, subjective individualism does not limit intentionality to the singular *form*, but restricts the class of possible *subjects*, or 'bearers', or 'owners' of intentions to single individuals.

What is the role of this distinction between two versions of 'Cartesian' individualism in the current debate? As mentioned above, Bratman and Searle both reject individualism in breaking away from the orthodox standard model of intentionality and, at the same time, resort to individualism when they see themselves confronted with what they perceive to be the ugly face of the group mind. In this apparently paradoxical venture, the distinction between the two versions of individualism comes in handy: both Bratman and Searle choose to depart from *one version of individualism* in setting apart their respective concepts of collective intentionality from the standard model, and to resort to the *other version of individualism* in order to banish the group mind. Interestingly, however, they do not seem to agree on which version of individualism to throw out, and which one to keep! Bratman's conception of shared intentionality seems to go beyond subjective individualism in some respects and to hold on to formal individualism, whereas Searle makes the opposite move. This results in a rather peculiar constellation: who is right? Or should it turn out that both are equally right (when they reject one form of individualism) and wrong (when they endorse the other form of individualism)?

Before turning to this question, let me establish the facts about Bratman's and Searle's respective forms of individualism. Bratman argues that what he calls "shared intention" or "group intention" is not anything single individuals can 'have' for themselves, but rather an "interrelation" (1999: 114) or an "interlocking web" (1999: 9) of what goes on in the minds of many individuals. Thus it seems clear that, in this account, "shared intentionality" cannot be structurally independent of external relations. We have to stand in actual relations for our intentionality to be shared. What makes our intentionality *shared* goes beyond the minds of single individuals. Thus Bratman seems to reject subjective individualism in the sense defined above. At the same time, however, Bratman deems it necessary to endorse formal individualism in order to not get stuck with the group mind. He hastens to declare that the relations presupposed in shared intentions are not tantamount to some "fusion" of individual agents to a "superagent" (1999: 111, 122ff.). In this respect, Bratman stresses that his account is thoroughly "reductive in spirit" (1999: 108). He rejects the idea that individuals literally *share* what they have in mind when intending together by emphasizing that the element of "we-ness" involved in what individuals intend when engaging in shared intentionality is reducible to a special

[19] Another term that is in use for this view is "methodological solipsism" (see Searle 1990). In the given context, I find this term misleading, for the question at stake here is clearly not simply a question concerning methodology, but an *ontological* question concerning the subject or bearer of intentionality.

form of I-intentionality: intentions of the form "I intend that we J" on the part of the individuals involved, together with mutual knowledge of this intentionality and some matching relation between what the individuals intend, make up shared intentionality.[20]

Significantly, it was not the interrelationistic move beyond individualism that has been most criticized in the debate of Bratman's account, but rather his reduction of we-intentions to sets of I-intentions (i.e. the move Bratman makes to avert the group mind). The upshot of a long discussion[21] is the following. It seems that I have to take myself to be in a rather *influential position* within the we-group in order to form intentions of the form "I intend that we J" do.[22] The reason is that intentions of this form extend into other people's intentional domains. Whoever has such intentionality in a sense intends other people's behavior. Thus he or she has to take others to be *responsive* to his or her own intentionality in a suitable way: he or she has to take his or her intentionality to be of some *influence* on other people's behavior (cf. remember the Principle of Intentional Self-Confidence as described above in chapter 1). Is this compatible with Bratman's account? My own impression (which I cannot argue for at length here) is that this "influence-condition" (1999: 116) shows that Bratman's account presupposes the element of sharedness it aims to explain.

Consider the following example. If we jointly intend to meet for lunch today, it does not seem necessary – indeed it is redundant – for me to form an intention of the form "I intend that we meet for lunch today" (rather, I will typically form some we-derivative [Sellars 1980: 99] or participatory [Kutz 2000a] intention of the kind "I intend to call you before noon to arrange a meeting place"). If and only if I take myself to be in a position to have a say in that matter, I might form an *additional* intention that *specifies* the content of our we-intention, and this additional intention might be of the form "I intend that we J_x" (e.g., "I intend that we have lunch together *at the Japanese restaurant*"). But intentions of this sort *presuppose* shared intentions instead of being their building blocks. It is only because *we intend J* that I can have intentions of the form "I intend that we J_x". Thus it seems that Bratman's "reductive" account of shared intentionality "in terms of attitudes and actions of the individuals involved" (1999: 108) simply fails to give an account of the crucial

[20] Cf. Bratman's conceptual analysis in Bratman (1999: 105).

[21] See Baier (1997a/b); Stoutland (1997, 2002); Velleman (1997); Bratman (1999: 149–156); Kutz (2000a/b). Concerning the question of whether or not intention should be put in propositional or action-referential terms cf. the remarks above in Chapter 1.

[22] At first glance, it might appear that intentions of the form "I intend that we J" are simply impossible. It is widely recognized that one cannot intend what one believes oneself to be incapable of doing (cf. Baier 1970: 658), and it seems clear that one cannot perform the actions of others (even though one can, of course, act on their behalf). Thus it seems to be impossible to include the actions of others in one's own intentions in the way it would be required in order to form intentions of the form "I intend that we J". Upon closer consideration, however, it seems that in these cases, one does not have to intend the actions of others in a straightforward sense, but that one simply has to take one's own intending to be of sufficient *influence* on the other participants so as to bring about their respective intentions to perform their part (Bratman 1999: 116).

element of collectiveness that is presupposed at its very base, because he endorses formal individualism.

In this respect, it seems to be Searle rather than Bratman who gets things right. For Searle stoutly opposes formal individualism. In his view, collective intentionality is a "primitive phenomenon" which is not to be reduced to any set of individual "I intends" plus mutual knowledge.[23] Yet Searle, too, sees himself confronted with the group mind, and he, too, resorts to individualism in order to banish it. However, in his conception, it is *subjective individualism* that plays that latter role. Searle argues that methodological solipsism is the only way to navigate safely between the two inacceptable alternatives, i.e. the Scylla of reductive formal individualism on the one hand and the Charybdis of the group mind on the other (the latter Searle calls "a perfectly dreadful metaphysical excrescence"[24]). Thus he claims, in a modern version of Descartes' *genius malignus* argument, that our collective intentionality is entirely in the heads of individuals and structurally independent of anything beyond individual minds.[25] Even a solitary brain in a vat that is somehow fed with the appropriate stimulus, or just lost in its dreams, and that is thus deluded about its real circumstances, could have intentions of the form "we intend". In Searle's view, the "we intend" (which is not reducible to individual "I intends") is something single individuals have in their minds, and this is structurally independent of whether or not these minds stand in actual relations to the world – or to each other, for that matter.

Many of Searle's critics think that this is wrong[26] – with good reason, I believe. It is true, of course, that the actual nonexistence of a group or the inexistence of co-members does not necessarily prevent individuals from intending *as if* they were members of that group. Just imagine the case of a dream about being one of the

---

[23] Among others, Searle puts forward the following two arguments against reductionism. First, common knowledge does not amount to the "sense of collectivity" involved in collective intending (1990). Second, our mind is too limited for the infinite iterations of knowledge implied in the "common knowledge" approach: "I think my poor brain will not carry that many beliefs" (Searle 1998b: 15).

[24] Searle (1998b: 150); see also Searle (1990: 404); Searle (1998a: 118).

[25] "Anything we say about collective intentionality must meet the following conditions of adequacy:
Constraint 1
It must be consistent with the fact that society consists of nothing but individuals. Since society consists entirely of individuals, there cannot be a group mind or group consciousness. All consciousness is in individual minds, in individual brains.
Constraint 2
It must be consistent with the fact that the structure of any individual's intentionality has to be independent of the fact of whether or not he is getting things right, whether or not he is radically mistaken about what is actually occurring. And this constraint applies as much to collective intentionality as it does to individual intentionality. One way to put this constraint is to say that the account must be consistent with the fact that all intentionality, whether collective or individual, could be had by a brain in a vat or by a set of brains in vats"; Searle (1990: 406ff.).

[26] Most forcefully, Anthonie W. M. Meijers has argued against the endorsement of methodological solipsism in the theory of collective intentionality; cf. Meijers (1994, 2002, 2003). See also Johansson (2003); Hornsby (1997); Waldenfels (1996); Celano (1999: esp. p. 239ff.); Turner (1999: 216, fn. 20).

dancers in the group in the first version of Henri Matisse's 'Dance'.[27] If we neglect the question about whether or not metaphysical socialism about intentionality is true, it seems obvious that minds that do not stand in actual relations to others, or brains in vats for that matter, may well *take themselves* to be members of a team. The decisive question, however, is whether or not even philosophers who accept subjective individualism concerning *individual* intentionality should be social externalists concerning *collective* intentionality (as I suggest they should): should we take such intentionality to be *collective intentionality* that just happens to be *mistaken* in some way, or shouldn't we rather say that this intentionality does not qualify as collective intentionality in the first place?

Searle advocates the first alternative. In his view, "the existence of collective intentionality does not imply the existence of collectives actually satisfying the content of that intentionality" (1997a: 450). For Searle, such cases as the one just mentioned simply show that "my presupposition that my intentionality is collective may be mistaken" (1990: 407). He admits that the case of a solitary brain in a vat having we-intentions constitutes a mistake of a very special kind[28] "which violates the Cartesian assumption that we cannot be mistaken about our intentions" (1998b: 150). But this "price to pay" (ibid.) seems all the more moderate since, in Searle's view, the Cartesian idea about the transparency of our intentionality proves to be wrong even in the case of individual intentionality, and is thus a mistaken notion anyway. Contrary to what Descartes thought, we can be mistaken about one of our intentional states[29] – why should this not be true for collective intentionality? In other words, the fact that there might be no actual collectivity involved in our collective intentionality boils down to just another way in which intentions can be mistaken – something that fails to touch the very *structure of our intentionality* itself.

Together with Searle's critics, I would like to put forward a different view. It seems to me that, by conceptually restricting collective intentionality to what is in individual minds, Searle misses a crucial element in the makeup of collective intentionality, which is the very element that Bratman emphasizes in his departure from solipsism and his move towards an interrelationalistic account of collective intentionality. For the sake of the argument, let's accept the *general* possibility of envatted brains[30] in order to take a closer look on Searle's claim that collective intentionality

---

[27] Cf. the reproduction on the cover of Searle's "Construction of Social Reality" (1995).

[28] What is in question here "is not simply a failure to achieve the conditions of satisfaction of an intentional state and is not simply a breakdown in the background"; cf. Searle (1990: 407).

[29] Cf. Searle (1998a: 69ff.). Here, Searle distinguishes four ways in which we can be mistaken about our consciousness in general, and our intentions in particular including self-deception (e.g. in the case of suppression of our dark sides) and misinterpretation (as in the case of somebody who takes his temporary infatuation to be real love).

[30] The possibility of 'envatted brains' is highly controversial. Putnam (1981) argues that it can be ruled out a priori; Dennett (1991) argues that the computational performance required in order to provide the 'envatted brain' with the appropriate input would be "computationally intractable on even the fastest computer". The question to be addressed here, however, is not whether or not 'envatted brains' are possible at all, but whether or not those brains, *if they were possible*, could be said to share intentions.

"could be had by a brain in a vat or by a set of brains in vats" (1990: 407). Imagine Ann and Beth visiting the Museum of Modern Art together; they happen to be the only visitors at the time. On the first floor, they get lost in the sight of the first version of Henri Matisse's 'Dance' (the example is a homage to Searle's [1995] cover illustration). Now a figure that is hard to avoid in envatted brains thought experiments puts in his appearance: an evil scientist creeps up behind our two heroines, and while Beth runs away screaming for help, he anesthetizes Ann for a minute, puts her brain in a vat and connects it to a computer that provides it with the appropriate input so that Ann has the impression of simply continuing to contemplate Matisse's 'Dance' together with Beth, just as if nothing had happened. Now it seems that, in her vat, Ann still has intentionality that conforms to Searle's concept of collective intentionality. All the intentionality Ann has in her mind seems to remain unchanged in subject, intentional mode, and content. It is still Ann's intentionality, and she still intends to contemplate Matisse's 'Dance' together with Beth (or, for that matter: she still intends her contemplating Matisse's 'Dance' as her 'we-derivative' individual contribution to her and Beth's shared intentional activity). Thus Ann may still have intentionality that is collective in *form* and that has 'collectivity' or 'sharedness' in its *content*. However, it is clear from the semantics of the verb "to share" alone that, in her vat, whatever she might *believe* she intends, Ann does not *in fact* share the intention to contemplate Matisse's 'Dance' together with Beth anymore. It is obvious (and trivially true) that the sharedness of intentionality is not a matter of the *form* or *content* of one single individual's intentionality alone. The question that turns out to be non-trivial is: what is it that has to be *added* to the picture for there to be proper shared intentionality?

In spite of its obvious importance to the theory of collective intentionality, Searle seems to be strangely disinterested in this question.[31] It seems clear, however, that within his internalist framework, the following answer imposes itself: When Ann and Beth were in fact *sharing* their intention to contemplate Matisse's dance together, they *both* (we-) intended to contemplate Matisse's dance (or to contemplate Matisse's 'Dance' individually as their contribution to their shared intentional activity). After the evil scientist's intervention, however, *only Ann* (we-) intends to contemplate Matisse's 'Dance' together with Beth. Beth, on her part, has no such intentionality any more, for she now intends to do something quite different, i.e. to run to the information desk of the Museum of Modern Art as quickly as she can to call for help. Thus it might seem that the answer to the question of what the intentionality Ann has in her vat lacks in order to qualify as *shared* intentionality can be found in Beth's head. In order for (we-) intentionality to be shared, *all* participants have to have the appropriate (we-) intentions, which is not the case anymore in the given situation.

This answer, however, is clearly deficient. Here is why. Imagine the story of Ann, Beth, and the evil scientist to continue as follows. After the evil scientist has finished his business with Ann, he goes after Beth. On the ground floor, halfway to the exit, he catches up with her, anesthetizes her and puts her brain in a separate vat,

[31] This has not escaped Bratman's notice (1999: 116, 145).

connecting it to a second computer. Beth forgets all that has happened since the evil scientist appeared on the scene, and she is provided with the appropriate input so that the intentionality she has is "We contemplate Matisse's 'Dance"' or "I contemplate Matisse's 'Dance' as my part of our contemplating". Now let's get back to Ann, who is still in her vat on the upper floor. According to the internalist-minded view of the sharedness of intentionality I just sketched above, it seems that Ann's intentionality has become *shared intentionality* again in the very moment when the evil scientist switched on Beth's computer. For now, just as before the evil scientist's intervention, *both* Ann and Beth have intentionality of the form "we are contemplating Matisse's 'Dance' together" or "I am contemplating Matisse's 'Dance' as my part of our shared contemplating". This conclusion, however, is implausible; intentionality does not become shared intentionality just because completely independently of each other, two brains just happen to have appropriately 'matching' illusions. If shared intentionality is not a matter of what goes on inside an individual head *alone*, it is not a matter of what goes on inside *different* heads, either. In order to find out about the sharedness of Ann's and Beth's intentionality, it is not enough to check only what is in the minds of the two individuals. As Anthonie W. M. Meijers has pointed out most forcefully, sharedness is a matter of the *relations* between minds, i.e. something that "transcend[s] the boundaries of [. . .] the 'brain in a vat"' (Meijers (1994: 7).

The further question is: what *kind* of relation is required for intentionality to be shared? What *sort* of 'connection' do we have to add to the Searlean picture of isolated minds for there to be proper sharedness? I cannot aspire to giving a straightforward answer here, but shall restrict myself to contrasting my ideas with Meijers's, whose critical discussion of Searle's account of collective intentionality I still believe to be the most important one in the existing literature.

Meijers opposes Searle's theory of collective intentionality in at least two ways. Firstly, he argues that Searle's internalism has to be given up in favor of a relational account. Secondly, Meijers criticizes Searle's view that collective intentionality does not involve social normativity in the form of commitments, obligations, and entitlements.[32] Along this line, Meijers argues that we have to give up Searle's *cognitivism* in favor of a *normativist* stance.[33] It seems that on Meijers's view, these two moves are internally connected, or even just two different aspects of one and the same move, so that the "radical relational approach" to collective intentionality he advocates somehow *has* to be a normativist one. This becomes clear from passages such as the following: "Cognitive attitudes are not sufficient to explain the *sharing* of intentionality. Normative attitudes have to be part of the analysis." Is Meijers right? And if not: why does he think the relations in question have to be normative ones?

In Meijers's view, collective intentionality "arises [. . .] out of the act of agreeing",[34] and it is within an analysis of this aspect of collective intentionality that

[32] In Searle's view, any such normative phenomena come into play only with the use of language, which is logically posterior to collective intentionality (see Searle (2001a: Chaps. 5 and 6)).

[33] For a detailed normativist account of shared intentionality see Gilbert (1996).

[34] Meijers (1994: 89); cf. ibid.: 104ff., 143.

we have to go beyond Searle's internalism and move towards a relational account (Meijers 2003: 176, 167). Applying this view to the above example, it is essential for the very sharedness of Ann and Beth's intention to contemplate Matisse's 'Dance' together that there is some kind of (implicit) *agreement* between them, some shared *commitment* to do so, which to some degree *obliges* Ann and Beth to do their part and at the same time *entitles* both of them to rebuke the other if she does not perform her part.[35] Meijers argues that Searle's internalist theory of collective intentionality cannot account for these normative aspects. It seems clear that in her vat, Ann still might *believe* there to be an (implicit) agreement between herself and Beth to contemplate the paintings on exhibit together; however, as Meijers points out, there is a difference between *agreeing* and *seeming to agree* (Meijers 2003: 179), and it is this difference that the Searlean approach to collective intentionality cannot account for because of its internalist limitations. Just looking at what goes on in the individual mind of Ann there is no telling whether she is in an *actual* agreement with Beth or just *believes* herself to be so. In the latter case, however, there is no agreement and thus no shared intentionality between Ann and Beth.

I believe that this argument is sound in itself, but I do not see why the difference it hinges on – the difference between "A-ing" and "seeming to A" – should be specific to the normative aspects or forms of shared intentionality. It seems to me that the same point can be made within a cognitivist view, according to which the relations involved in collective intentional states do not necessarily have to be of the normative sort. Consider the following case. Imagine Ann and Beth to be dyed-in-the-wool Searleans. For them, their visit to the museum does not involve any kind of commitment, obligation, or entitlement whatsoever. However strange this might seem, it just happens to conform to their usual practice that any of them may walk away from the common enterprise at any time, without owing the other any further explanation. There is no agreement whatsoever between them; they are both just regular visitors to the museum on Sunday afternoons who over time have come to see their individual visits as part of a common enterprise. The first to come usually waits at the entrance for the other; if, as it sometimes happens, the other does not show up, she does not feel that the missing party has wronged her, or that she is entitled to some explanation. On the face of it at least, the intentionality involved is thus strictly limited to cognitive aspects. My point is the following: Even though there is no agreement, entitlement or obligation around, it still makes a difference if Ann just *believes she shares* the intention to visit the museum together or if she *actually shares* this intention. If Searle cannot account for the normative aspects of shared intentionality within his internalist framework, he cannot account for the purely cognitive aspects either.[36] Thus it does not seem necessary to connect the two

[35] For a normativist account of shared intentionality see Gilbert (1996).

[36] Indeed it seems that there is a great deal of purely cognitive components involved in shared intentionality. Take the case of Anne and Beth in their respective vats. What sort of connection has to be established between them in order for them to share intentions? It seems that a great deal of delusion is compatible with shared intentionality. Indeed there is even a sense in which the two brains in vats might said to be share their intentions, if their respective sources of input

issues Meijers raises against Searle. One does not *have* to take a normativist stance on collective intentionality in order to follow Meijers's advice to give up Searle's internalism in favor of a "radical relational approach" (Meijers 2003: 167).

§8 Collective Intentionality: Irreducible and Relational

The comparison between Bratman's and Searle's account reveals complementary strengths and weaknesses. On the one hand, Searle is right in renouncing formal individualism which seems to be the weakness of Bratman's account. Shared intentionality is not reducible to sets of I-intentions, because the I-intentions individuals *have* when taking part in a shared activity *presuppose* shared intentionality. On the other hand, Bratman is right in departing from subjective individualism. As Bratman makes clear (and contrary to what seems to be a consequence of Searle's approach), it is only *in relations* that individuals share intentions. Thus it seems that Bratman and Searle, in their respective departures from the Cartesian model of intentionality, both get stuck half-way in the project of developing an adequate account of collective intentionality. Their problem is that they let go of only one version of individualism, while holding on to the other. An adequate account of collective intentionality, however, has to depart from the Cartesian individualistic picture of intentionality not just in renouncing *either* formal *or* subjective individualism. It has to be *both* non-reductive *and* relational.

By way of giving a rough outline of my idea of such an account, I should like to propose two tentative theses, concerning the relational (i) and irreducible (ii) character of collective intentionality, respectively.

i. Social normativity arises out of shared intentionality (and not the other way around)

Agreement-based accounts of shared intentionality beg the question because any sort of agreement *presupposes* shared intentionality. The act of agreeing is itself a move *within* a shared intentional activity (whereas not all cases of shared intentional activity involve agreement). Shared intentions which are based on agreement do, of course, exist. But these are shared intentions of a special (and especially complex) kind. Therefore, it seems that they should not be taken to be the 'paradigm case' of an analysis of shared intentionality. Thus I agree with Searle (as well as with Raimo and Maj Tuomela; cf. 2003b) that collective intentions do not *by themselves* involve social normativity in the form of obligations and entitlements. At the same time, however, I find the Searlean picture of completely normativity-free collective intentional activities (such as the one depicted above) rather askew. If the sharedness of intentionality is not necessarily *in itself* socially normative, it has *socially normative*

---

are appropriately connected (so as to make Anne believe that Beth does what Beth believes she does, and conversely). "The Matrix" provides a vivid illustration of shared cooperative activity by appropriately interconnected "brains in vats".

consequences. As was pointed out repeatedly in sociological theory, proper social norms arise out of merely habitual social practices such as customs (Geiger 1987). It seems to be almost inconceivable that we might engage in shared intentional activities over an extended period of time without our *cognitive* expectations concerning the actions of others gradually turning into *normative* expectations (which entails no less than a reversal of the direction of fit of the respective attitude; I shall come back to this in the concluding chapter of this book). My conjecture is that these socially normative *consequences* of shared intentions stem from the pre-socially normative (or, in Tuomela's words, from an "instrumentally 'normative"') *implication* of *any* kind of shared intention. The implication in question is the following. To the individuals involved, a shared intention provides a *reason* to form an appropriate personal intention (i.e. the intention to perform one's part).[37] In a pre-socially normative sense, if we intend, I *ought* to do my part in what we intend. This *normative* relation between shared intention and individual we-intention, however, does not exclude the possibility of overriding contrary reasons or simple weakness of will. Thus it seems possible that *we intend x* without me *intending to do my individual part* (even without my having a pro-attitude towards our shared aim; cf. Chapter 3 below). If this perspective on the relation between shared intentions and personal 'contributive intentions' is correct, it has far-reaching consequences: the analysis of shared intentionality cannot be based on an analysis of what individuals personally intend when sharing an intention, but the analysis of what individuals intend when taking part in a shared intentional activity has to be based on an analysis of the structure of shared intentionality (Tuomela 2002b). Or, to use Edmund Husserl's concept of *foundation* (Rota 1989):

ii. Shared intentionality is the foundation of individual (we-) intentionality (and not the other way around)

The German phenomenologist Gerda Walther, thinking about shared intentionality in the early 1920s of the last century, and struggling against her own individualistic preconceptions, which she had taken over from Husserlian phenomenology, used a striking metaphor for what seems to be at stake in the move towards a non-reductive and relational account. She talked of a "Copernican Turn" (Walther 1923: 98) from an analysis of sharedness that is derived from an analysis of the intentionality of the participating individuals to an understanding of the intentionality of the participating individuals that is based on a solid concept of sharedness. But how should

[37] This seems to be at odds with John Broome's (2001) claim that intentions are not, as such, reasons for action. Broome's point is that if I have no justifying reason to intend to A, but intend to A (perhaps because I am mistaken about the relevant facts), I would have reason to do what I have no reason to intend to do, which does not seem plausible. In the current case, however, the issue at stake is not the relation between intention and action, but between collective and individual (participatory) intention. And as far as contributive action is concerned, it seems obvious that in normatively stabilized cases of joint action there are normative expectations from the other participant's side involved. Even if there is no justifying reason for us to intend J, I might have a reason to do my part; insofar as under some description, our aim is to do what we have reason to do, however, my part may well be to voice my doubts about there being a reason. For a more detailed discussion cf. Chapter 3 below.

such a turn be possible without simply replacing the individual with the collective as the source and bearer of intentionality? As seen above, fear of the group mind plays an important role in driving some of the most important accounts of shared intentionality back into the seemingly safe harbor of individualism. Thus it seems important to address the question: is this fear justified? Will an account that neither embraces formal individualism nor subjective individualism end up getting stuck with the group mind?

I believe that any such reservations against a non-individualist (i.e. non-reductive and relational) account of collective intentionality are mistaken. As seen, the whole trouble with the group mind arises from the attempt to give some acceptable answer to the question: who is the subject that *has* collective intentions? To whom can this intentionality be attributed as its source, bearer, or owner? And this question, innocent as it might look, is heavily loaded with historical ballast that we should, I think, simply jettison and leave behind.

Only in the last decades, have we successfully managed to get rid of Descartes' quest for absolute certainty in philosophy.[38] However, the Cartesian preoccupation with the "subject" still persists. It is still a deeply rooted idea that where there is intentionality there has to be a somebody who "has" it as its owner, source, or bearer.[39] It is the fact that most philosophers of collective intentionality hold on to this assumption that gives rise to the fear that by moving too far away from individualism, we are running the risk of getting stuck with the group mind. Yet there is a simple way out of the individualistic dilemma – or double-bind – in which current collective intentionality analysis seems to have gotten stuck: it consists in overcoming the "Cartesian Brainwash" by ceasing to address the "who has it" question. Collective intentions are not intentions of the kind anybody *has* – not single individuals, and not some super-agent. For collective intentionality is not subjective. It is relational. Collective intentionality is an intentionality which people *share*.

---

[38] Even Searle, who is by some accused of sticking to the Cartesian "epistemological" paradigm in philosophy (cf., e.g., Dreyfus 1993), says explicitly that he is not "a part of the Cartesian tradition of trying to overcome skepticism and provide a secure foundation for knowledge"; Searle (2001b: 173).

[39] The preoccupation with the subject or "bearer" of intentionality seems to stem from what is perhaps Descartes' most durable insight. I myself have a privileged position among all the things I might be acquainted with. However deluded I am about the world – and, we can add, about my intentions – there seems to be something incorrigible or infallible involved in my self-awareness. Even if I live in complete delusion about all my beliefs, there is still something that I simply cannot get wrong: it is in fact *myself* whom I am aware of when reflecting on my beliefs and desires. Even if some madness has me in its tightest grip, misleading me into thinking that I am Henri Matisse, it is still infallibly *me myself* whom I take to be Henri Matisse – it is not, for example, the actual Henri Matisse whom I take to be Henri Matisse. This insight seems to be at the base of Descartes' claim that what is really certain and indubitable about my thinking is the subject, the bearer of intentionality, i.e. the thinking "I". Now it seems obvious that, however right this might be concerning the "I" of individual intentions, it does not apply to the "we" of collective intentions, for I might easily be mistaken in any collective belief or intention.

In conclusion, I should get back to the initial ontological question concerning the relation between the individual and the collective. I think that with the illusion of the group mind the urge to drive actual collectivity out of the concept of collective intentionality vanishes, too. A theory of collective intentionality that is both non-reductive *and* relational does not require any logical or ontological primacy for the aims, attitudes and emotions of the individuals over the actual existence of the group. This does not mean, however, and conversely, that it requires the logical or ontological primacy of the group over the individual. Searle seems to think that we have to make our choice between these two versions of the Philosophy of Society: either we put the we-intentionality of individuals or the collective itself first (qua "ontological primitive" [Searle 1997a: 449] that somehow precedes our we-intending). His choice, then, is the first alternative: "Collective intentionality gives rise to the collective and not the other way around" (ibid.). It seems to me that, conceived of like this, the whole question about the relation between the individual and society is wrongly put. It implies what I should like to contest: that collective intentionality and actual collectivity are two different things. Only because, in the current debate, collectivity was driven out of the concept of collective intentionality in the first place does the question about how one is related to the other arise. If collective intentionality is not subjective, but relational, there is no need to postulate any ontological order of hierarchy between the analysis of collective intentionality and the ontology of groups. Because, in a relational sense, collective intentionality is what the ontology of groups is all about.

What is the bearing of this result on the Philosophy of Society? It seems that overcoming the 'Cartesian Brainwash' means to break away from the individualistic approach to Philosophy of Society, and to move towards a more heterodox view. In light of a post-Cartesian concept of collective intentionality, it appears that the orthodox slogan that "there is no society, only individuals who interact with each other"[40] is not outright wrong, but simply meaningless. Most forms of interaction involve collective intentionality, and collective intentionality is what society in the most basic meaning of the word *is*.

[40] Cf. Elster (1989: 259) quoting Margaret Thatcher. It should not be forgotten, however, that Thatcher continued as follows: "there are only individuals, *and there are families*" (Woman's Own Magazine, 10/3/1987).

Chapter 3
On Not Doing One's Part

Dissidence and the Normativity of Collective Intention

Looking at the illustrations and examples that are usually chosen to discuss the structure of joint action in the literature on collective intentionality, one might sometimes feel magically disburdened of the troubles of mundane interactions and transferred into an idyll of smooth cooperation. Thus John Searle, in his writings on the topic, treats us to some homey scenes, with people dexterously preparing a meal or playing music together (Searle 1990). Out on the street, Raimo Tuomela has a traveling choir joining their forces to push their broken-down bus up the hill (Tuomela 1995: 137–138). In the park nearby, some of Robert Sugden's team-thinkers are engaged in a game of football (Sugden 2000), while beneath the trees, Margaret Gilbert's committed couple of friends are out on their Sunday afternoon walk (Gilbert 1996). Michael Bratman, finally, takes us farther out into the woods. Here, we meet Abe and Barbara, who are patiently and diligently working together to pump water to their weekend cottage (Bratman 1999: 150–151).

As has already been noted (Baier 1997b), this is a world of cooperation between keen and capable contributors, in which the exceptions to the rule are few indeed. Thus there is a disabled person in Tuomela's group of car pushers. But even he is doing his best to promote the collective venture by making encouraging remarks from his wheelchair (Tuomela 1995: 138; Tuomela 1991: 272ff.). With all these willing volunteers and keen contributors, the world to which our philosophers of collective intentionality invite us is an attractive place. At the same time, however, it might appear to be somewhat unreal. Even if we accept such limitations as the concentration on small-scale cooperation among adults, something is missing. Where have all those negligent, sloppy, unfocused, forgetful and weak-willed people gone whom we know from real life, all those who for some reason or another fail to do (or even fail to intend to do) their part in their collective projects? And where are those recalcitrant, fractious and unruly fellows who not only fail to do their part, but even have the intention *not to do their part*, in a shared cooperative activity?

These people and their role in collective intentionality analysis will be the focus of the following considerations. I shall argue that their absence in the received accounts has led to a skewed view of collective intentionality. This chapter is in three parts. In the first section, some remarks concerning the possibility and limits of not doing one's part shall be made, and some conjectures will be offered as to why the phenomenon in question is widely left out in the received view on collective inten-

H.B. Schmid, *Plural Action: Essays in Philosophy and Social Science*, Contributions to Phenomenology 58, © Springer Science+Business Media B.V. 2009

tionality and shared cooperative activity (§9). The second part addresses the question of how our view of the structure of collective intentionality should be changed in order to make room for this phenomenon (§10). The third section is on a special (and especially important) kind of 'not doing one's part', i.e. the case of dissidence (§11). Here, as already in §2, I shall focus on Raimo and Maj Tuomela's analysis of the structure and role of dissident participation. Some concluding remarks sum up the considerations made in this chapter and open up a perspective on social ontology.

§9 Joint Intention and Individual Participation

The bias towards smooth cooperation that is found in much of the literature on collective intentionality is hardly just a matter of the examples chosen. Rather, it seems to be a direct consequence of a very basic feature of how collective intentionality and collective intentional activity are approached. The emphasis on active participation naturally results from the fact that, according to most of the received accounts, the intentionality of actively participating individuals is what collective intentionality analysis is all about. Indeed, this might seem to be quite natural as a starting point, for how could a collective intention (or even a shared intentional activity) ever come into being without single individuals committing themselves to doing their share? And how could individuals commit themselves in this way without forming an intention to act accordingly? What other than some form of conditional personal commitment of the form "I will if you will" (cf. Gilbert 1989) should be at the origin of collective intentions? And how should a collective intention, once it is formed, persist over time, let alone become effective, without the corresponding participatory intentions in place? Thus Raimo Tuomela puts forth an account in which the participating individuals' intentions to do their part, together with a mutual belief concerning the other participants' intention, play the central role (cf. Tuomela and Miller 1988). Similarly, the building blocks of Michael Bratman's analysis of shared intentional activity are individual intentions of the form "I intend that we J", together with some common knowledge of these individual intentions (Bratman 1999: 105). In John Searle's view, too, the intentionality of the participating individuals is the focus of the analysis. In the Searlean version of the story, the participating individuals have some irreducible intention of the we-form kind in their mind (Searle 1995: 26). Regardless of the considerable differences between these accounts, the common thread is that whatever collective intentionality might be, it is somehow built of (or supervenes on) the keen contributors' and eager participants' intentionality.

In the following, I shall use the label *participation theory of collective intentionality* in order to refer to this common feature of the received accounts of collective intention. It is true, though, that the participatory element is a matter of degree. One can find accounts that seem to require much less intentional activity, commitment and dedication to the common cause from the part of the participating individuals than the abovementioned accounts seem to do. But even in the cooperatively least demanding account I know of (Kutz 2000a), collective intentional activity is ultimately based in some "participatory intention" of the individuals involved.

In the following, I shall try to present some rather tentative considerations, which are meant to cast some doubt on the participatory view. It seems to me that the participation theory of collective intentionality is at odds with some strands of our intuitive, pre-theoretic understanding of what it means to share an intention. There seem to be everyday cases where in ordinary language we refer to collective actions or collective intentional states without implying any appropriate individual contribution or contributive intentions, or any individual we-intending. Such examples have even been discussed in the literature. Thus Annette Baier quotes the case of some member of a family gathering, sitting on the porch of the venue where the gathering is held. Asked about what she is doing, she replies "*we* are dancing a reel, but I quickly had enough of it, so I am sitting out" (Baier 1997b: 26; my emphasis). From the perspective of ordinary language at least, Baier's non-dancer appears to be fully justified in using the word "we" rather than "they" to refer to the dancers, even though, obviously, she herself does not participate in the joint intentional activity any more, and may not intend to participate again in the future. How do such examples square with the intuition that lies behind the participatory approach? How do we identify those justified to use the term "we" from those not justified to do so if not by means of their appropriate contributive mental states? After all, even Baier will have to admit that the "we" in question isn't open to anyone. So what is it that distinguishes those belonging to the "we" in question from the outsiders? Baier's example seems to suggest that the "we" of the person's statement is not the group of dancers really, but rather the family, of which she is a member independently of whether or not she takes part in the dancing. Thus it would seem that individuals are justified in using "we" with reference to a group jointly doing (or intending to do) x without participating (or having the intention to participate) if and only if their use of "we" is justified in some other context, i.e. *independently* of the collective intention or joint activity in question. Upon closer look, however, one might find that this need not necessarily be the case. Imagine the dancing in Baier's example to be the joint intentional activity of a spontaneous group rather than that of a family. In the course of hours-long intense performance of social dances, all participants have come to see themselves as members of this spontaneous group of dancers; there is, however, no "we-ness" beyond the context of the ongoing activity. Even now, it does not seem obvious why some member that decided to stop and chill out on the porch should not refer to the ongoing dancing activity in the 'we'-form. In this case, the identity criteria of the group are set by the history of the social practice itself rather than by external factors.

Baier seems to think that such apparently *nonparticipatory uses of 'we'*, as one might call them, are possible only in collective activities involving larger groups, where the shared intentional activity in question does not immediately break down just because one single individual decides to opt out (Baier 1997b: 26). It seems, however, that something similar is true of smaller group (and even dyadic) activities. Consider the following case. The two of us are walking across campus to have lunch at the cafeteria. After a short while, we find ourselves in the middle of a lively discussion, and without stopping to walk, we are more and more drawn into the topic of our debate, forgetting all about the world around us. After a while, I ask you, as if

waking from a dream: "*What in the world are we doing here?*" It would seem quite natural if you were to reply: "*We are going to the cafeteria together!*", even though you know that there is no suitable participatory intentionality on my part involved here, since I do not even seem to be aware of what we are doing. How, then, is this everyday language use of 'we' justified? One possibility is to give it a *performative* reading.[1] Sometimes the word "we" creates its own conditions of propriety. In Margaret Gilbert's sense, your use of "we" would then have to be considered *initiatory* (Gilbert 1989: 178–179) rather than *constative*, turning the half-conscious coordinated behavior that was going on during the time of my oblivion back into a case of proper shared intentional activity again. While this might well be the case, however, it seems obvious that there are legitimate uses of "we" in this situation which are not initiatory, or performative. Just consider the third-person equivalent. Even before your use of "we", a distant (but appropriately informed) observer of the scene could veridically report that *we* were walking to the cafeteria together, and that I simply had forgotten what we were doing. How does this second example square with the participatory view? One way to make this case fit the standard view seems to be to say that I *did* intend to do my part all along, even though I had forgotten about this intention temporarily. My participatory intentionality was there all along – unconscious intentionality, as it were. This claim is not as strange as it might first appear. Not all intentional states need to be *conscious* states (cf. Searle 1983: chap. 1). Thus the fact that I did not know what we were doing, in itself, does not seem to prove that I didn't intend to do my part. But this does not pull the example's teeth, as we might modify it so as to exclude the case of unconscious participatory intentionality. There is a difference between the case where I am *currently unaware* of my intention (such as in the case where I intend to walk to the cafeteria with you, even though I am not conscious of this intention right now, because my focus is fully on our discussion), on the one hand, and the case where I *have forgotten about* and *am unable to recall* my intentions (such as in the case of our example), on the other hand. Whereas the former case poses no problem for the participatory view, it seems that the extension of 'unconscious intentional states' to the latter case simply overstretches this concept. It is true that one need not be currently *aware* of one's intentions. One might argue, however, that these intentions should at least be *accessible* under appropriate circumstances. It seems to fit nicely with this line of thought that, in ordinary language, we would probably be more likely to refer to my participatory intention (i.e. my intending to go to the cafeteria with you) as something that simply *ceased to exist* at the point where I was not just concerned with our discussion, but *forgot* all about the whole project, rather than to treat it as something that continues to exists in some unconscious form.

The tentative and rather sketchy character of these considerations notwithstanding, it appears that, in a pre-theoretic sense at least, it is true of the two of us that we indeed *collectively intend* to go to the cafeteria together, even though, under the given circumstances, you might be alone in actually having the appropriate

[1] I am grateful to Frank Hindriks for pointing this out to me.

intentionality. Once again, the result is that, as compared to the standards of the participation theory of collective intentionality, the strict participation requirements have to be relaxed in order to accommodate such intuitive cases.

Similar cases can be found for the other types collective intentional states, i.e. cognitive and affective attitudes. In a variation of an example originally put forward by Edith Stein in a short passage on shared emotions (Stein 1922: 120), let us consider the case of some professional association's official gathering, during which the president expresses "our" deeply felt grief at the sudden death of a senior honorary member. Imagine that some overly critical investigative journalist wanted to check out the truth of this statement. He arranges for a short private conversation with each participant, asking her or him whether she or he really felt grief (or a similar emotion) when the president made his statement. As it turns out, all but a few answer in the negative. Some felt sympathy with the bereaved, others simply felt dissatisfaction with the president's poor performance as a speaker. Hardly anyone actually felt any grief. The question is: should we think that this little survey proves the president's statement wrong? I think we should not; if anything is wrong here it is the journalist's understanding of the truth conditions of the statement in question. References to such shared emotions do not, for their truth, depend upon what the majority of the individuals in question actually happen to feel. "Our" grief is a collective intentional state, but it does not seem to be of the simple participatory kind. In this case, as in the other cases mentioned above, it simply seems to be *misleading* to approach the collective intention in question from the side of the intentions of the participating individuals.

Admittedly, these considerations are of an explorative character and do not extend beyond the phenomenological level of the description of pre-theoretic intuitions. It seems, however, that if we accept at least some of these phenomenological descriptions as adequate to our pre-theoretic intuitions, and if we take these pre-theoretical intuitions to be a touchstone for our theories, the conceptual requirements of individual participation for the existence of collective intentions have to be relaxed. Shared intentionality is compatible with much more individual non-contribution – much more *deviance* or *dissidence*, as it were – than the participatory theory of collective intentionality suggests. This is the line Raimo and Maj Tuomela have taken when they set out to liberalize the participation (and "pro-attitude" requirements) so as to make room for dissidents in their account of acting as a group member (cf. e.g. Tuomela and Tuomela 2003a: 15). The aim of the following chapters is to contribute to this a discussion of some of the issues involved in this move from a phenomenological point of view.

§10 Participation and Normativity

One might think that the abovementioned examples for seemingly 'non-participatory' collective intentions are unfit for the analysis of the structure of collective intentionality, because they are non-standard cases. There is something

special, *extreme*, or even to some degree *pathological* about most if not all of these cases. When, in the first of the abovementioned examples, Baier's non-participant on the porch uses "we" to refer to the group of dancers, she seems to owe us an explanation. Asked about what *she* is doing, she cannot just say "we are dancing a reel inside" without further comment. What she needs to explain is why, then, she is not taking part herself. Indeed, Baier herself has her adding "but I quickly had enough of it, so I am sitting out" (Baier 1997b: 26). Obviously, that explanation would not have been needed if she had either referred to the dancers in the third person plural form, or taken an active part in the dancing. This shows that our non-participatory case is a *special, non-standard* one; and something similar seems to be true for all of the abovementioned examples.

The fact that non-participatory cases are non-standard does *not* mean, however, that such cases are *marginal* or *irrelevant* for the study and analysis of collective intentionality. Here, as in so many other examples, it is in the light of *non-standard* cases that normality reveals its basic traits. These cases confront us with the decisive question: *why*, precisely, is it that in the deviant, non-participatory cases some extra explanation is needed? The most obvious answer is: when people share an intention, they can be *expected* to have the appropriate participatory intentionality. If *we* are dancing, I can be expected to intend to do my part. "Expectation" is meant not in a purely cognitive, but in a *normative* sense (cf. Chapter 10 below), meaning that we do not cease to expect people to take part in their collective activities just because in some cases, such as the ones mentioned above, these expectations are not met. Baier's non-dancer on the porch needs to explain her behavior, because if she refers to the dancing as something "we" do, she can be *expected* to be taking part herself, which she is not.

All of the abovementioned cases seem to be conceptually possible. This goes against the participation-based approach to collective intentionality. However, it has to be admitted that something is not as it *ought* to be with these non-participatory cases. If we are going to the cafeteria, I *ought* to intend my walking to the cafeteria as a part of our going there together. *Mutatis mutandis*, the same is true for the other cases. This is why the other participants as well as external observers of the scene will *expect* me to do my part. If this is true, it seems to follow that the relation between the participating individuals' intentionality and the collective intention is not a *constitutive*, but a *normative* relation. If, in the "normal" case, those involved in a shared intentional activity have the appropriate intentionality, this is not because, in some sense, these participatory intentions are what collective intentions are "made of", or because they (in some loose sense of the word) *constitute* collective intentions, or because collective intentions *supervene* on individual intentions, but because the individuals involved in shared intentional activities *ought* to (intend to) do their part. To put it bluntly: the stuff collective intentions are made of is *normativity*.

This raises further questions. *In what sense* are collective intentions normative? In the existing literature, there is a controversy between *normativist* and *non-normativist* accounts. Margaret Gilbert (Gilbert 1989, 1996) and Anthonie Meijers (Meijers 1994: 89, 104ff.) claim that collective intentional activities are based

on (tacit) agreement and always involve *obligations* and *entitlements*, whereas John Searle, Michael Bratman (1999), Raimo and Maj Tuomela (Toumela and Tuomela 2003a) and Robert Sugden 2000, 2003) hold that participation in shared intentional activities does not *per se* involve obligations and entitlements. My claim concerning the normative character of the relation between collective intentions and the participatory intentions of the individuals notwithstanding, I take the *non-normativist* side in this controversy. According to the view developed in the previous chapter, non-normativist accounts are right in emphasizing that, in principle, shared intentional activities are possible without there being any *proper obligations* and entitlements involved. At the same time, however, I argued against the anti-normativists. If the sharedness of intentionality is not necessarily *in itself* socially normative, it inevitably has *socially normative consequences*. Social norms arise out of merely habitual social practices such as customs. It is inconceivable to engage in shared intentional activities over an extended period of time without our *cognitive* expectations concerning the actions of others gradually turning into *normative* expectations. It is true that people might agree not to develop normative expectations concerning each other's contributions; but if this were done, it would itself be part of the normative infrastructure of the joint intentional activity in question. I conjecture that these socially normative *consequences* of shared intentions stem from a *pre-socially* normative (or, in Tuomela's words, from an "instrumentally 'normative"') *implication* of *any* kind of shared intention: To the individuals involved, a collective intention provides a *reason* to form an appropriate personal intention (i.e. the intention to perform one's part, or to we-intend the collective activity). Contrary to what the existing normativist accounts suggest, the basic sense in which one *ought* to (intend to) do one's part is *not* that of *social* normativity (duty or obligation to (we-)intend x), but of pre-social normativity (having a *reason* to (we-)intend x). In a pre-socially normative sense, I *ought* to do my part in what we intend, and any obligation or entitlement that might come to play a part in shared intentional activity ultimately arises from this pre-social normativity.

This *pre-socially normative* relation between shared intention and individual we-intention, however, does not exclude the possibility of overriding contrary reasons or simple weakness of will (we shall come back to this in §13). Thus it is possible that *we intend x* without *me intending to do my part of x* (or without me we-intending x, or even without me having any pro-attitude towards our shared aim at all). This explains both why the abovementioned cases of shared intentional activities are conceptually possible, and why they are *non-standard* (and to some degree *deficient)* cases that need further explanation.[2]

[2] This version of normativism avoids some circularity that seems to be characteristic of the existing normativist accounts of shared intentional activity. Gilbert seems to hold that obligations and entitlements are essential to shared intentional activity because any shared intentional activity ultimately originates in some form of (perhaps tacit) *agreement*. However, for something to count as an (however tacit) agreement, some form of shared intentional activity has already to be in place, for "agreeing" is not anything single individuals can do, but something people have to do *together*. Thus shared intentional activities cannot be based in agreements, because agreements are nothing but a special kind of shared intentional phenomena themselves.

It might appear, however, that the use of the term "reason", as applied to collective intention, and the talk of "ought" concerning individual participation is mistaken. John Broome has argued forcefully that we should not take any kind of normativity to be a matter of having reasons. Broome (2000) distinguishes between reasons and normative requirements, and he argues (2001) that intentions are not reasons. The difference between reasons and normative requirements become most obvious in the case of conflict. Sorting out conflicts between reasons is a simple matter of weighing. Reasons "add" to our decisions according to their "weight", and they do so even when they are outweighed by contrary reasons. This is different with requirements. Reasons are "slack", requirements are "strict": requirements do not allow for degrees: either P is required to A or she is not. Requirements cannot be outweighed. This does not mean that one always has a *reason* to do what one is required to do. On the contrary, one might well have a reason *not* to do so, for example, if a given goal requires one to employ morally unacceptable means. In that case, one has a reason to *give up* the requirement. Thus, in the case of conflicting normative requirements, there must be something wrong in a way that has no equivalent in the case of conflicting reasons.

Broome argues that the normativity of intention is a matter of normative requirements rather than a matter of reasons because P cannot conclude that she *should* act in terms of having a *reason* to act on an intention just because she happens to have that intention; for it might well be that she does not have a reason for having that intention – or for acting on it – after all. In itself, intention is not reason-providing.

So much for Broome's view. He does not comment on the structure of the normativity of collective intention. At first sight, the normativity of collective intention seems to be made of the same cloth. It is plausible to assume that cases of conflicts between collective intentions and personal intentions are more than just a matter of weighing conflicting reasons. There really is something "abnormal" about those conflicts, which seems to point at the fact that the normativity of collective intentions, just as the normativity of individual intention, is a matter of normative requirements, where the connection between the premises and the conclusion is tight. At the same time, however, collective intentions differ from individual intentions in ways that resemble reasons more than normative requirements. Individual intentions break down if there is constant conflict with practical conclusions; in the collective case, the connection is less tight. Moreover, collective intentions usually involve normative expectations concerning one's behavior from the other participant's parts; these act on one's practical conclusions in the way of reasons rather than in terms of normative requirements. Thus there is a sense in which collective intentions are reason-providing in a way in which individual intentions are not.[3]

[3] I am grateful to Juliette Gloor for suggesting this to me.

§11 The Structure of Dissidence

Focusing on the normative nature of collective intentional phenomena, it is tempting to concentrate exclusively on the 'top down dependency' between collective intentions and the intentionality of the individuals. This, however, might lead to an overly conventionalist view of collective intentionality phenomena, and indeed it is only half of the story. For the normative dependency goes both ways, i.e. not just from the collective to the individual level, but from the individual to the collective level, too. That "our" intentions provide a stance from which I can critically assess my own intentions and actions (Rosenberg 1980: 159) is not sufficient for establishing this relation. It is important not to neglect the fact that the converse is also true. Where the intentionality and actions of the participants do not correspond to the collective intentions and collectively intended actions, this might not just be the individuals' fault, as a conventionalist view of collective intentionality has it. In a sense, one can say that the collective intentions *ought* to correspond to what the individuals intend or do. Just as individuals can critically assess their individual intentions and actions from the standpoint of collective intentions, practices, and projects, these intentions, practices, and projects need to be critically assessed in light of what the individual participants intend or do.

This brings us to a very peculiar and especially important way of not doing one's part: the case of *dissidence*. In the narrow sense, the term has been mostly used to refer those oppositional intellectuals who, perhaps protected by their international reputation, were to some degree tolerated by the eastern European and Soviet communist regimes (for an account of the ethos of those dissidents cf. Tucker 2001). It seems to me that at least two features are essential to a wider concept of dissidence. The first is that dissidents are group members with a different idea of what "our" collective plans, projects, and actions should be. In this sense, dissidents are basically *dissenters*, who play an important role in any kind of communal practice (cf. Sunstein 2003). However, dissidence is not just about dissent, and not all dissent is dissident dissent. Dissidence is different from more common kinds of opposition, and is indeed an important part of any process of collective intention formation, and is even institutionalized in any democratic process in larger groups. Whereas, in expressing their oppositional views, *dissenters* simply do their part in a collective procedure of collective intentionality formation, *dissidents* find themselves to some degree *outside* of the collectively accepted communal practices and institutional frameworks. This brings us back to the topic of this chapter. Dissidents see themselves forced *not to do their part* in our communal practices in order to do justice to their views of what the collective should be. In short, to be a dissident means more than just to have a differing view about our communal plans and projects. It means to refrain from participating in some sort of communal practice, too.

Let us take a closer look at the interplay of these two core features of dissidence. It seems that the element of dissent and the refusal to do one's part can play different roles in different forms of dissidence. Without doubt, in the paradigmatic case of dissidence, the dissident's refusal to do her or his part in some communal practice is simply a direct consequence of her or his dissenting view. Consider the case of

David Henry Thoreau's "civil disobedience", his refusal to pay his poll tax because of his strong disapproval of the Mexican war (Thoreau [1849] 1967), or Martin Luther's famous "Here I stand, I can do no other". However, history and literature are full of dissidents of a very different and much less intellectualist kind, which I find theoretically more challenging than such principled and well-considered disobedience. Quite often, the refusal to do one's part is not a proper premeditated and principled behavior that is derived from the dissident's dissenting views, but a more or less spontaneous act. Thus, in Friedrich Schiller's play, William Tell walks by the pole with the tyrant's hat on top without bowing, not because he rejects this humiliating practice of symbolic submission, but simply because he happens to be talking to his son and does not pay attention.[4] Or, to quote a real life example: when on December 1, 1955, in Montgomery, Alabama, Rosa Parks refused to give up her bus seat to a white man and earned herself the title of the mother of the civil rights movement, this was not meant as a gesture against race segregation. As Parks recalled the events later, she was simply weary and tired after a full day's work as a tailor's assistant, and her body ached (Raines 1977: 40–43). In these and similar cases, the dissident's deviant behavior is a spontaneous failure to conform to the communal practices rather than a principled and premeditated act.[5] Thus one could distinguish a kind of 'principled dissidence' from this more spontaneous kind, in which the failure to do one's part plays much more important a role than that of a more or less symbolic gesture and direct consequence of one's differing views.

Annette Baier has pointed out that in the realm of the social, "the power of the negative is an important power, and our dissidents and awkward customers, our [. . .] cultural subversives provide that power" (Baier 1997: 37). Especially in the latter, spontaneous case of dissidence, this power flows not just from the dissenting views concerning communal affairs, but from the irritation that ultimately stems from more or less spontaneous failures to do one's part in normative communal practices. And such cases show how important *not doing one's part* can be even as a source of inspiration and a starting point for a whole renewal and reconstruction of communal intentions, projects, and practices.

The tendency to underrate the importance of the role of non-participation with respect to shared action seems to be so deeply seated that it even extends to some accounts that take a thoroughly positive stance towards dissidence. An especially significant example of this tendency is Karol Wojtyla's "Introduction to participation" in the appendix to his book on the Acting Person (Wojtyla 1979: 323ff.).

---

[4] Later on in this decisive scene, Tell even apologizes to the tyrant, pointing out that he did not act on purpose when he failed to greet the tyrant's hat ("Verzeiht mir, lieber Herr! Aus Unbedacht, nicht aus Verachtung Eurer ist's geschehn"). It is quite revealing, however, that in the libretto to Gioacchino Rossini's famous opera, this entire scene is altered to fit the *more conventional* view of principled, premeditated heroism. Here, the tyrant's guards order Tell to bow to the tyrant's hat, which he explicitly refuses to do (act III, scene III).

[5] It has been repeatedly pointed out that contrary to her "official" image as a quiet and observant citizen, Rosa Parks was active in the civil rights movement long before the Montgomery bus incident (Sparks 1997). The fact that she was well aware of the injustice of segregation laws, however, does not, in itself, prove false her own account of the events.

In this phenomenologically-minded analysis, the structure of shared action and its relation to the concept of a person is examined. According to Wojtyla, the essence of participation is *solidarity*, which is described as "constant readiness to accept and to realize one's share in the community" in accordance with existing collective practices (Wojtyla 1979: 341). However, Wojtyla is far from underestimating the importance of opposition and resistance. Writing only shortly before the insurrection of the "Solidarnosc" workers' movement against the communist rule in his homeland of Poland, Wojtyla (the later pope John Paul II) makes clear that solidarity is not incompatible with opposition: "those who stand up in opposition do not intend thereby to cut themselves off from their community. On the contrary, they seek their own place within the community" (Wojtyla 1979: 343). In contrast to this praise of the attitude of opposition, however, Wojtyla identifies two forms of "denial of participation", which he criticizes for being "inauthentic": avoidance and conformism (Wojtyla 1979: 346ff.). Avoidance amounts to "a lack of participation and in being absent from the community." By contrast to this, conformism is "a mere semblance of participation, a superficial compliance which lacks conviction and authentic engagement" (Wojtyla 1979: 346–347). It seems, however, that here, as in so many other cases, the inclusion of non-participation into the concept of shared action stops half-way. It should be extended to those attitudes that Wojtyla criticizes for being *inauthentic*. For the power of change and renewal does not come only from those who *voice their dissent*, but from those who simply *refuse to participate* (and perhaps choose the "exit" option rather than the "voice" option [Hirschman 1970]), too. Our admiration for the courage and bravery of those who speak up and fight within their communities is not diminished if we pay due respect to the important role of the refugees, the boat people, and all those who under unbearable circumstances simply withdraw from participation.

A structurally similar critical point can be made with respect to the second of Wojtyla's "inauthentic attitudes". Conformism in terms of "a mere semblance of participation, a superficial compliance which lacks conviction and authentic engagement" can be a very effective means to bring about change and renewal. The practice of *working to rule* might serve as an example. Working to rule is indeed a retraction of "authentic engagement", of effort and cooperation, a withdrawal of those aspects of work that go beyond the formal regulations, but are nevertheless necessary for the efficient and profitable functioning of an industry or an administration. Thus by working to rule, pressure can be exerted. In circumstances where communal practices are based on the member's willingness to do more than just 'their part' in terms of their formal duties – and this is true to some degree in all organizations – conformism may sometimes be more effective in subverting an existing collective practice than any open opposition.

The positive account of the attitude of dissidence that is sometimes given in the sociological literature is deceiving.[6] If dissidence and the other abovementioned

[6] Thus the French sociologist Michel Maffesoli, quoting Pierre Sansot, concludes his short examination of the *dynamics of dissidence* with an overly optimistic statement that is characteristic of his entire analysis, saying that "la dissidence [. . .] a le 'don de desserrer extraordinairement l'étreinte des forces de répression"' (Maffesoli 1978: 111).

ways of not doing one's part are virtues, these virtues are only of *secondary* status. Their value depends on the moral quality of the communal practice against which they are directed. Such ethical questions, however, need not be of our concern here. Instead, I shall conclude with a short recapitulation of the main point of this chapter, and with a remark concerning the ontology of groups.

In an earlier account of the structure of teamwork, the following claim can be found: "if a team has a goal p, then each member has p as an individual goal" (Levesque and Cohen 1991: 499). In other words, individuals not having p as their goal find themselves excluded from the team *on conceptual grounds* by this participatory account of shared goals. The same is true for all variants of the participation theory of collective intentionality, at least insofar as groups are taken to exist by virtue of collective intentionality (cf. Searle 1997a: 449). This seems to be at odds with our commonsensical, pre-theoretic understanding of collectively intentional phenomena. We often refer to collective intentions, where there is a great deal of *intentional deviance* on the part of the individuals in question. Our everyday use of "we" seems to show that one does not lose one's membership simply because one fails to intend to do one's part. Within his admirably detailed account of collective intentionality, Raimo Tuomela was the first to address this issue, proposing the concept of "non-operative membership" (Tuomela 1991: 272ff.) to resolve this problem. Tuomela requires his non-operative members to show some pro-attitude towards the collective venture. Here, the non-operative members, too, have to do their part in the collective venture, in order to qualify as members of the team. In his more recent analysis of group responsibility, however, Raimo and Maj Tuomela go one step further by including a thoroughly dissident sense of acting as a group member. I think that this is indeed the path that should be followed. And in my view, one main step in loosening the conceptual constraints on group membership and acting as a group member should be to switch from "is" to "ought" in the analysis of the relation between collective intentions and the intentionality of the participating individuals. It seems clear that the team members *ought* to have the appropriate intentionality when the team intends x. This, however, does not rule out their failure to live up to this normative requirement, as can be observed in actual cases. A normativist account of the relation between collective intentionality and the intentionality of the participating individuals, such as was proposed above, is, however, in direct conflict with the standard participation theory. That in a pre-social sense, one *ought* to do one's part in what we intend *presupposes* what is ruled out by the participation theory. To say that one *ought* to (intend to) do one's part makes sense only if it is at least *conceptually possible* that one does not in fact (intend to) do one's part. To get back to the above case: if a team has goal p, the members *ought* to have p as their individual goal (in a weak instrumental sense of "ought"). This, however, *presupposes* the *conceptual* possibility that the team has goal p without the single individuals having the goals they *ought* to have.[7]

[7] Note that this need not be a real possibility. One does not have to have the actual option to do otherwise in order to be normatively required to do one's part. Deviance need only be conceivable, or conceptually possible. Thus this normativist account does not rule out a compatibilist view.

Chapter 4
Shared Feelings

Towards a Phenomenology of Collective Affective Intentionality

Up until very recently, collective intentionality analysis has almost exclusively been concerned with shared intentions and – more recently – with shared beliefs. Next to no effort, however, has been devoted to the analysis of collective *affective* intentionality so far. This apparent research lacuna seems all the more surprising because the analysis of emotions has been among the key topics of international philosophical research for the last 20 years, at least, with a plethora of conferences, monographs and collected volumes on the topic. Thus the question is: why is the emotional dimension so conspicuously absent from collective intentionality?

§12 Affective Intentionality: A Matter of Feelings

One might first think that this lacuna simply reflects the general neglect of the affective in the wider history of the theory of intentionality. In the Anglo-Saxon context, intentionality has always primarily been regarded as a matter of *conative* (or practical) intentionality, whereas on the continental side, cognitive intentionality has usually served as the paradigmatic case. Of course, there are many exceptions to this rule on both sides of the "great divide" (among these Carl Stumpf, Martin Heidegger, Max Scheler, Gilbert Ryle). But these exceptions cannot distract from the rule: in neither of these two traditions of intentionality analysis has the affective received an amount of attention that comes anywhere close to the one devoted to the cognitive and conative types of intentionality. Most philosophers of intentionality simply did not deem the emotional important enough to deserve particular attention. Thus it might indeed seem that the neglect of the affective dimension in the debate on collective intentionality is part of a more general anti-affective strand, which runs through most of the received philosophy of intentionality.

Something of a mirror image of this can be seen in early conceptual analyses of the affective. Just as the affective has been neglected in the theory of intentionality, intentionality has been neglected in the theory of the affective in turn. According to what has come to be called "feeling theories" of the affective (Solomon 2003, 2006), the intentional is at best of minor conceptual importance to the affective. The essential feature of the affective is seen in the *feelings* qua states of *bodily arousal*, as

H.B. Schmid, *Plural Action: Essays in Philosophy and Social Science*, Contributions to Phenomenology 58, © Springer Science+Business Media B.V. 2009

in William James' *What is an Emotion* (1884). As a consequence, the question of the intentionality of emotional states is not seen as particularly important. In this view, a theory of the affective should be concerned with a taxonomy of feeling experiences, and with the analysis of the causal role of states of arousal, rather than with such ventures as intentional analysis. Thus the emerging picture is as follows. There is a rift between an affectivity-free theory of intentionality on the one hand, and an intentionality-free view of the affective on the other. Let us call this constellation the intentionality/feeling dichotomy.

At first sight, it might seem that much time must have passed since the days of this dichotomy. In contrast to the days of the "old" image of the emotions, current philosophy of the emotions tends to be particularly interested in the very feature that has so shamefully been neglected in much of earlier philosophy. The following has come to be widely accepted: it is central to our understanding of the emotions that they be seen as a way our mind is "directed at" the world, and how our world is "given" to us. In short, emotions are now acknowledged as *intentional states* with specific content, rather than as mere states of bodily arousal.

But this is only part of the story. Upon closer examination it becomes obvious that, with the recent turn to the intentionality of emotions, the days of the intentionality/feeling dichotomy are definitively over. The view of the intentionality of the emotions that has been dominant in most of the recent debate tracks both the neglect of the emotional in much of the earlier theory of intentionality, and the fact that the intentionality of the emotions has been overlooked in earlier theories of the affective, to one and the same source. Both shortcomings, it is claimed, are due to the fact that the emotions have been wrongly identified with the *phenomenal aspects* of the emotions, i.e. with the *feelings*. In many theories of the intentionality of emotions, this is regarded as a mistake. Turning away from feeling theories, cognitivists claim that emotions are intentional in somewhat the traditional sense of intentionality. Emotions are intentional *insofar – and only insofar – as they imply cognitive and conative aspects*, i.e. *beliefs and action dispositions*. And this does not seem implausible at all. Most certainly, the belief that one is in some kind of danger does play a role in standard cases of fear, as does the tendency to flee and seek refuge, and similarly for all other emotions. The "directedness" of emotions – the question of what fear, anger, joy, pride are all about – is clearly a matter of those very beliefs and action tendencies, rather than of the feelings involved in standard cases of emotions. Perhaps the strongest arguments for this latter claim are the following. First, "feelings do not have 'directions'" (Solomon [1975] 2003: 4): no fact about the twinge we feel, taken in itself, can tell us whether the twinge is indeed one of, say, remorse, rather than one of rheumatism. Thus it seems that whoever takes the twinge to be the core feature of emotional experience cannot at the same time focus on affective intentionality. Insofar as such feelings describe the phenomenological aspects of emotions, it seems that the phenomenological approach to the emotions misses the intentional character of our affective lives. Second, as opposed to emotions, feelings cannot be true or false (Solomon 2006: Lect. 3). Third, feelings are by definition conscious states; emotions, on the other hand, can be unconscious, as in the famous case in which after years of psychoanalysis, one finds out about the

previously unconscious emotions one has had all along. Fourth, the emphasis on the "feeling" component tends to contribute to the "Cartesian" internalist image of the emotions, according to which emotions are "in the mind" rather than a form of engagement and entanglement with the world.

We shall come back to these issues shortly. For the moment, let us simply take the cognitivist view of the intentionality of emotions for granted, and ask how this view answers our initial question. How does this contribute to explaining why so little work has been devoted to the analysis of collective affective intentionality? If the story I have told so far is basically correct, one does not have to assume some particular anti-affective bias in order to explain the apparent neglect of shared affective intentionality in current collective intentionality analysis.[1] The explanation might actually be much simpler. If the cognitivist view of the intentionality of emotions is indeed right, the answer to the question concerning the reasons for the alleged research lacuna is simply that there is no such research lacuna. On this view, it simply *makes no sense* to weigh the amount of work done on collective conative and cognitive intentionality on the one hand against the amount of work devoted to collective affective intentionality on the other. To claim that the philosophers of collective intentionality have neglected the analysis of collective affective intentionality by focusing on the cognitive and conative dimensions of intentionality would amount to a simple *category mistake* – because in the cognitivist view, *shared intentions and shared beliefs are what the intentionality of shared emotions is all about*. Thus the analysis of shared emotions appears to be nothing more than another realm to which the tools of the received analysis of shared conative and cognitive intentionality can be applied.[2]

The decisive question concerning the alleged research lacuna therefore is the following: is the cognitivist view of the intentionality of the emotions right? In the light of the most recent developments in the philosophy of the emotions, it appears that there is no straightforward answer to this question. Most certainly, the cognitivist view is right in emphasizing that emotions do indeed imply some sorts of cognitive components – be it beliefs, judgments, perceptions, or other sorts of commitments.

---

[1] Yet, there may actually be such a tendency at work in some accounts of collective intentionality. Defending an anti-affective stance would be to claim that collective intentionality extends to shared beliefs and shared intentions, but not (or not to the same degree) to shared emotions. Among the received accounts of collective intentionality, Christopher Kutz (2000) seems to come closest to this view (cf. e.g. Kutz 2000: 196).

[2] This seems to be the case in Margaret Gilbert's very important contributions, which mark the most conspicuous exception to the general neglect of shared emotions in received collective intentionality analysis (cf. Gilbert 1997; Gilbert 2002). Gilbert discusses the case of collective guilt feelings. In standard cases, however, guilt presupposes action, as far as one tends to feel guilty for what one has done (or failed to do). This makes it particularly tempting to adopt a cognitivist view for the analysis of this particular shared emotion, and this is indeed what Gilbert does. Gilbert's analysis of shared guilt feelings uses the very tools she has developed for the analysis of joint action, particularly the notion of collective commitment (Gilbert 2002). Thus shared emotions appear to figure in as just another application of the theory. I think that, in order to avoid any potential bias, cooperatively less marked examples should be chosen. In the following, the case of shared grief is used as a paradigm. In contrast to guilt, grief does not imply action.

And many emotions are intrinsically related to action. At the same time, however, it has become increasingly clear over the last couple of years that the cognitivist view of the intentionality of emotions is distorted. In many respects, the intentionality of emotions is very *different* from the intentionality of beliefs or other forms of cognitive attitudes. This is obvious from the following facts: first, we can have emotions even in spite of contrary beliefs (this is the case in "emotional recalcitrance"; think of David Hume's example of the fear a person feels at the sight of the precipitate even when she is protected by an iron cage [Hume (1739/1740) 2000, book 1, part 3, sect. 13, §10], or of Charles Darwin's [1872: 38] similar experience with the snake behind the glass). It is implausible to explain these cases with the assumption that they entail having an inconsistent set of beliefs, because these cases do not seem to make us *irrational* in the same way inconsistent beliefs do.[3] Secondly, the respective relation of beliefs and emotions to the will is fundamentally different: We can neither believe nor have affective attitudes at will, but we can (and indeed very often do) cultivate our affective attitudes towards things in a way that has no direct analogy in the case of beliefs (cf. again Goldie 2000 for a detailed discussion). Thus in this respect, too, the intentionality of emotions is very different from that of ordinary beliefs.

If affective intentionality is different from the intentionality of ordinary beliefs (or other forms of cognitive intentionality), one might be tempted to think that this is because it is more like the intentionality of *desires* (or conative intentionality). This, again, might not seem implausible at first; being afraid implies the desire to flee and seek refuge, and similar action tendencies play an important role in many emotions. But this view does not hold up against closer scrutiny either. In fact, some emotions do not seem to involve *any* desire or action tendency at all, such as the pride one feels at the success of one's daughter; these emotions rather seem to be of perception-like quality (Goldie 2000). Thus it seems that the cognitivist account of emotions, instead of defending the importance of affective intentionality against ignorant feeling theorists, leads to a very distorted view of the intentionality of emotions. Cognitivism about emotions tends to model affective intentionality much too closely on the traditional model of intentionality of beliefs and desires, thereby neglecting the very peculiar way in which emotions are (or can be) intentional, and have cognitive content.

The intentionality of emotions does not seem to conform to the core feature of the belief/desire model of intentionality, which is the idea of a "direction of fit". Put very bluntly: whereas beliefs are supposed to fit the world, the world is supposed to fit our desires. In a core sense, however, it seems that affective intentionality has neither of these two "directions of fit." If emotions are indeed a form of "engagement" and

[3] In contrast to such cases of emotional recalcitrance, cognitivists naturally tend to emphasize those cases where emotions vanish as a consequence to the formation of some new beliefs. To use Robert Solomon's example, it seems natural to cease to be angry with James if one learns that it was not him who stole one's car after all. This is a very mature reaction; I'm afraid, however, that in the normal real-life case, the anger will quickly resurge under a new mask. Although the "normative judgments" may change, the emotions are likely to persist.

"entanglement with the world", as Robert Solomon emphasizes, this entanglement is not of the cognitivist kind. John Searle, to whom we owe the conceptual elaboration of the idea of a "direction of fit", sometimes speaks of the "nil direction of fit" with regard to the emotions – which is not a direction after all, as one might be tempted to add. It is typical of affective intentionality that the idea of a direction of fit is not particularly helpful to the analysis of these states.

Another difference between ordinary cognitive attitudes and affective attitudes is that the content of basic affective attitudes is non-propositional. Fear is *of* something, but its object need not be (part of) a proposition.

If affective intentionality is not just a matter of ordinary cognitive and conative attitudes – what is it, then? In light of the most recent developments in the philosophy of the emotions, it seems that we have to *overcome the intentionality/feeling dichotomy* in order to understand the way in which emotions are intentional. It is the deep-seated preconception that *feelings are not intentional* – a view that feeling theorists and their cognitivist opponents have held alike – which prevents us from seeing how emotions "disclose" the world.[4] By playing the intentionality of emotions off against feelings, the cognitivists drive a wedge between the intentionality of emotions on the one hand, and the phenomenology on the other, leaving the feelings the role of mere contingent *accompaniments* of emotions.[5]

Suffice it to say that the tide in the current debate has turned: more often than not, feelings are now being seen as a core feature of emotions. Many authors claim that feelings can be (and indeed are) intentional, that the intentionality of feelings is different from cognitive and conative intentionality, and that *feelings* are at the heart of affective intentionality (which does, of course, not mean that *all* feelings are affective (or emotional), as there are *cognitive* or *epistemic* feelings, too, such as the feeling of clarity). There are replies to all the cognitivist points mentioned above.[6] Feelings *do* have "directions", feelings *can* be appropriate or inappropriate (if not true or false), they are not just contingent "accompaniments" of emotions; the very notion of "unconscious emotion" (i.e. emotion without feeling) is skewed, if not outright mistaken; and if the emphasis on the role of feelings for the emotions is taken to convey some illicit 'Cartesian' internalism, this is a question of how feelings are conceived of rather than of the concept of feeling itself.

---

[4] For an early critique of feeling theories cf. Bedford (2003[1956/57]); the tendency to play the intentionality of emotions off against feeling theories is particularly conspicuous in Robert Solomon's work. Cf. among many other examples Solomon 1993.

[5] This view is also adopted by Margaret Gilbert in her analysis of collective guilt feelings (Gilbert 2002). Gilbert made it her business to defend the claim that in a certain sense, collectives can "feel guilt". Her analysis, however, makes it clear that the second part of her title – "collective guilt feelings" – is a misnomer. Gilbert does indeed claim that collectives can be said to "respond affectively" in terms of guilt, but she emphasizes that qua phenomenological ingredients, "feelings" are no necessary part of this emotion, but "only frequent concomitants" (Gilbert 2002: 119) or "accompaniment" (ibid.: 141) of emotions. Against this, Burleigh Wilkins has convincingly argued that a collective commitment without feeling might not be enough for there to be genuine collective guilt (Wilkins 2002).

[6] For points 1, 2 and 4 cf. Goldie (2000), or, for an even more radical view, Ratcliffe (2005); for point 3 cf. e.g. Clore (1994); LeDoux (1994).

One might call this recent development the "phenomenological turn in the philosophy of emotions". Here, feelings are conceived of as *intentional*– as "feelings towards", to use Peter Goldie's expression.[7] For the purpose of this chapter, I take this development for granted, and I shall address the question: what are the consequences of this phenomenological turn for the purpose of an analysis of *collective* affective intentionality? For it seems clear that if affective intentionality is a matter of *feelings* rather than just of beliefs and desires, the question of how affective intentionality can be collective cannot simply be answered by pointing towards the received accounts of shared beliefs and joint intentions. If the phenomenological view of the intentionality of emotions is right, then there *is* a research lacuna. And it seems plausible to assume that, on the basic level, it has to be filled with an analysis of *how feelings can be shared*. I shall proceed as follows. In the first step, I shall examine the structure of "feelings towards", and say a word on what it is about them that might be said to be shared (§13). I then turn to defending the metaphysical claim that, when people genuinely share a feeling, there is a sort of phenomenological fusion between the consciousness of the participating individuals (§§14–15).

§13 Shared Feelings: Content, Mode, and Subject

Let us first have a look at the intentionality of feelings. As any intentional state, feelings have a mode, a content, and a subject. The *mode* of a feeling defines the feeling as the *kind* of feeling it is. Thus, the feeling of fear of something is to be distinguished from the feeling of anger at the same object insofar as in fear it is felt to be dangerous, whereas in anger it is felt to be offensive. "Danger" and "offense" are the *formal objects* of the respective feelings and thus account for the *mode* of the feeling in question. Feelings of a particular kind come in different intensities and perhaps even qualities. Thus the feeling of joy can vary between wild exuberance and silent satisfaction.

As to the *content*, I suggest following Bennett Helm (2008) in drawing a distinction between the *target* and *focus* of a feeling. Put simply, the *target* is the object towards which the feeling is directed. In the case of fear of a dog, the target is the dog. The *focus*, by contrast is the object in the background of the feeling which is related to the target in such a way as to *make intelligible*, or *rationalize*, the *mode* of the feeling. If you encounter a big stray dog while jogging in the park, the focus of your fear of the dog will typically be yourself. It is because the dog is dangerous *to you* that your feeling towards the dog is one of fear. But one might also experience a

[7] In focusing on the phenomenological aspects of affective intentionality, however, one should be careful not to run feelings and emotions too closely together. Feelings extend to all sorts of things. One can "feel convinced", for example, or "feel real", or "feel stared at". There are many feelings which are not emotional in a narrow sense of the word/term. In the following, I shall use the term "affective intentionality" in a looser sense, extending it to those feelings which are intentional, but not strictly emotional.

feeling of fear towards a stray dog while observing it, from a distance, approaching a group of children. In that case, the target of the feeling is still the dog, but its focus is now on the children, and it is by virtue of their being threatened by the dog's presence that sense can be made of the fear.

For a focus-target relation to rationalize the mode of a feeling, however, there has to be an additional feature in place: the subject has to have some *concern* that serves to make the relation between focus and target *relevant* to the subject. If a person simply doesn't *care* about her own well-being, or about the safety of children, the fact that a dog might attack her, or the children, does not rationalize her feeling of fear. Insofar as they involve a concern, feelings are an indicator of what *matters* to us.

Concerns differ from mere *inclinations* in that they involve *patterns of emotional dispositions*. To be concerned about something means to feel afraid when that something is *threatened*, or to feel joy when it is *thriving*, or to feel sadness or grief when it is *lost*. Thus our concerns *structure our lives* in allowing us and others to make sense of our attitudes. Concerns can be *deeper* and *less deep* in accordance with just how much of our lives they structure.

This brings us to the *subject* of feelings. I take it that in a very basic sense of the word, our deepest concerns determine who we are. Selfhood and identity depend on our concerns. I do not wish to claim that concerns are per se worthwhile having; I assume that in many cases they are not. And I do not wish to claim that, in order to have a concern, an individual needs to have some knowledge of his or her concerns. In fact, I assume that people are often ignorant of – or even mistaken about – their concerns. But there is a very basic way in which our concerns, and thus our identities, are *indicated* by our feelings. Feelings are the light in which we see ourselves. To experience a feeling is to conceive of ourselves in terms of the underlying concern. Our identities as a friend, as a professional, as a lover of art are settled by affective attitudes. I propose to call the self-concept implicit in a feeling its *phenomenal subject*, and to distinguish it from the *ontic subject*. The ontic subject is the individual who has the feeling. The phenomenal subject, by contrast, is determined by the way in which the subject implicitly conceives of him- or herself *in* the feeling. The ontic subject answers the "who has it" question; the phenomenal subject answers the question *as who* the ontic subject has the feeling he or she has.

With these conceptual tools in hand, let us now turn to the question of collective emotions. There are three ways in which the question can be approached: via the mode, via the content and via the subject. I will briefly discuss each of these and claim that we need to combine these approaches.

(a) *Sharing the mode*. Let us start with the feature that settles the question of the *kind* of feeling at stake. It seems plausible to say that, in order to share a feeling, people have to experience a feeling of the *same* kind. But experiencing feelings of the same kind is clearly not sufficient. There is a sense in which we might say that the feeling of fear of dogs is widely shared in a given population, but this way of speaking borders on the metaphorical. For a feeling to be *genuinely* shared one person's being in an affective state of a certain mode cannot be entirely *independent* of the other person's being in the same affective mode. There have to be *connections*

of some sort. One way in which a person's affective state may account for another person's being in a state of the same kind is *affective contagion*. The example that is often given is the feeling of fear spreading in a group of children. But it is of course true (and has often been remarked) that affective contagion per se does not mean that there is anything shared about those affective states.

A somewhat richer structure is *affective attunement*. The basic idea goes back to Adam Smith's *Theory of Moral Sentiments* ([1759] 2000), and it has recently been revived by Robert Sugden (2003). Affective attunement involves at least three features: first, some capacity to read other people's mind, second, some sort of a preference for *affective conformity*, and third, a capacity to exert some *control* over the way one feels. The claim is that people *enjoy* being in the same affective state as those around them, independently of the mode of the feeling at stake. That's why sharing increases the joy, but diminishes the pain. Also, people aren't entirely *passive* concerning their feelings: there is some emotion regulation. People exercise their control over their feelings according to their preference for affective attunement. This mechanism seems to play an important role in explaining how *social norms of propriety for affective states*, or "feeling rules" (Hochschild and Arlie 1979), come about. By means of habitualization, people's expectations concerning the "affective attunement" to be reached in typical situations come to be normativized. The result is what I propose to call an *affective agreement*, which is a generally shared idea about the level of affective attunement expected to be reached in a given situation. People now have an idea of how one *ought to feel* under the given circumstances, and they aim at cultivating their own feelings so as to live up to those standards, and expect others to do likewise so as to maximize affective attunement.

(b) *Sharing the content*. Without doubt, affective attunement is an important phenomenon. But one might doubt whether it is *necessary* and *sufficient* for an affective state to be shared. For the first case, consider Max Scheler's example:

> Father and mother are standing at the dead body of a beloved child. They feel 'the same' grief, 'the same' pain. This does not mean: A experiences this grief and B experiences it, too, and in addition to that they know that they feel it – rather, their feeling is a feeling-together (*Mit-einanderfühlen*) (Scheler [1913] 1974: 23–24; my translation).

There needn't be any *interaction* or even *intercognition* between the two for their feeling to be genuinely shared. In this case, the sharing isn't a matter of contagion, mutual awareness, or affective attunement. It is, it seems, simply a matter of the *content* of the feeling. And affective attunement isn't *sufficient* for collective affective intentional states because where the content *isn't* shared, affective attunement cannot turn individual feelings into shared ones. Just imagine a group of schoolchildren playing hooky. Assume that each child is worried about his or her parent's reaction when the matter comes to light, as it certainly will. The children might be enjoying their affective attunement in their fear, and even reach some affective agreement, but there is a sense in which these children's fear isn't genuinely shared, because the children's feelings have different targets: each child is afraid of *his or her own* parent's reaction. So we might require of *genuinely* shared feelings that the target

of the participant's feeling be the same. But this isn't enough either. Imagine two siblings being afraid of their parent's sanction. The target is now the same; but if each of the siblings' worry is exclusively about the consequences for *him- or herself*, the feeling isn't really *shared*. Thus not only the target, but the focus, too, needs to be the same for a feeling to be shared.

This might seem intuitively plausible, but I argue that the sharing of target and focus isn't a *necessary* condition for an affective state to be collective (even though I leave open the possibility that, given mutual knowledge, it might be sufficient). The case of a shared feeling with a *different* topic and focus that I would like to consider is the culminating scene of the epic reflection on emotion that marks the beginning of western culture. King Priam's encounter with Achilles at the end of the Iliad is perhaps still one of the greatest scenes of an affective meeting of minds in world literature. The story goes like this. After 10 years of battle, Achilles has slain Hector in revenge for his friend Patroclus. Yet Achilles' thirst for revenge turns out to be insatiable; he goes on mutilating Hector's body, much to the Trojan's and even the Olympians' distress. After many outrages, the Olympians decide that something needs to be done about Achilles' feelings. Achilles has to be brought back into *emotional tune* with the basic world order. How can this be done? On Zeus' advice, and at great risk to his life, King Priam sneaks into the Greeks' camp in order to plead with Achilles for his son's body. Here is what Priam says to Achilles, and Achilles' reaction in Homer's words:

> Respect the gods, Achilles, and take pity on me, remembering your own father. I am more piteous far than he, and have endured what no other mortal on the face of earth has yet endured, to reach out my hand to the face of the man who has slain my sons.' So he spoke, and in Achilles he roused desire to weep for his father; and he took the old man by the hand, and gently pushed him away from him. So the two remembered – the one remembered man-slaying Hector and wept loudly, collapsed at Achilles' feet, but Achilles wept for his own father, and now again for Patroclus; and the sound of their moaning went up through the house. (Iliad, Book 24, 503–512)

Immediately after that, and with the help of a nice little ransom, Achilles decides to hand over Hector's body to Priam, so that world order is finally restored. With Hector's burial, the Iliad, which started with Achilles' fury which separates him from his community, comes to a conclusion. Achilles is brought back into emotional tune with the world around him.

The decisive question about that scene is: How does Achilles' grief for his father's abandonement combine with Priam's grief for the loss of Hector so as to move Achilles to an act of goodwill towards Priam? I believe the answer to be that there is something in his own grief that Achilles recognizes in Priam's. He recognizes that the feeling is *shared*. Yet it is clear from Homer's description that neither *target* nor *focus* of their feelings are shared. The target of Priam's grief is Hector, and the focus is Hector or himself, or perhaps Troy; the target and presumably the focus of Achilles' grief, by contrast, is his father Pelleas and Patroclus. The fascinating thing about Priam and Achilles' emotional encounter is that the target and focus of their feelings seem to be more suited to setting the two up against each other, rather than to allowing for any affective meeting of minds. After all, it is

partly because of the Trojans that Achilles has abandoned his father, and is going to die soon, and it is because of Achilles that Hector is dead. Thus target and focus are anything but shared. If the feeling of grief connects the two, it is rather by means of the shared *concern* behind the target-focus relation. Priam's grief for his son combines with Achilles' grief for his father's abandonment so as to move Achilles to an act of goodwill towards Priam *because Achilles recognizes his own concern with Pelleas' being deprived of Achilles in Priam's grief for the loss of Hector*. In order to do so, however, Achilles has to move from Pelleas to fathers *in general*. This involves reconceiving of himself *as a son* rather than as Achilles, and that means a shift in the phenomenal subject of his affective attitude. This fits nicely with the usual interpretation that is given of the Iliad, according to which the whole epos is about Achilles' affective withdrawal from his community in wrath in books 1–17, his acting out of purely *individual feelings* in books 18–23, and his finally being able to feel *as a human being* again in his sympathy with Priam in book 24.

(c) *Sharing the subject*. This brings us to the final component to be considered when thinking about collective feelings: the *subject*. Margaret Gilbert (2002b) has argued that for an affective state to be genuinely collective, there has to be a way to ascribe the emotion in question to a *collective* rather than just to individuals. Collective feelings have a *plural* rather than a *singular* subject. I think that this basic idea is right, but that her account suffers from the way she conceives of the plural subjecthood of the affective states in question. Gilbert sees plural subjects as "command centers" (2000: 5) *of their own* which are constituted by the participant's joint commitment. Thus, in Gilbert's view, there seems to be a sense in which collective emotions *cannot* be the participant's. But, as I have already argued, for there to be emotions, there have to be qualitative states; and there seem to be no qualitative states over and above the consciousness of individuals. There is no "what it is like" for a group apart from the individual's experience.[8] So it seems that insofar as collective emotions are qualitative states, they cannot be attributed to collectives and thereby be genuine collective emotions.

The alternative I would like to propose is to conceive of the plural subjecthood of shared affective states in terms of *phenomenal subjectivity* while holding on to the *ontic* claim that only individuals can have emotions in the full, qualitatively rich sense of the word. For an affective state to be collective, people have to share a concern. Sharing a concern leads them to *identify* with each other, or with the group, by conceiving of themselves, as part of the feeling, in terms of a collective identity. Feelings can be ascribed to groups by virtue of their member's experiencing their feelings *as* members of the group.

[8] Cf., e.g., Pettit 2003; however, Knobe and Prinz (2007) have discovered that people are more reluctant to say that collectives *feel* emotions than that collectives *have* emotions. However, this is much less markedly so in eastern cultures, and even in the west, the level of assent to such ascriptions isn't as low as one would expect. Needless to say, this supports the argument developed in this paper.

§14 Individualism About Feelings

With this result, let me now turn to a metaphysical question. It is this: in what sense of the word are feelings to be genuinely *shared?* What does the term *sharing* really mean in this context? Let us have a look at the simplest and most basic use of the word. Consider the case in which I propose to share a bottle of wine with you. Certainly, I do not thereby suggest that you and I each open a bottle, the two bottles being of the same vintage, or brand. Rather, I suggest that we enjoy *one and the same (token) bottle*. Similarly, the idea of sharing a car is not that of each one driving his or her own car, the cars being of the same brand. The point is to use *one and the same (token)* car *together*. The idea is this: one car, many users, one cake, many pieces, one apartment, many inhabitants, and so on, and so forth. This is what I will call the straightforward sense of sharing. In the straightforward sense, sharing is not a matter of type, or of qualitative identity (i.e. of having different things that are somehow similar), but a matter of token, or *numerical identity*. This straightforward sense of the concept of sharing, however, does not seem to apply in many of the cases of sharing a feeling discussed above. Note that in contrast to the straightforward sense, it is part and parcel of the very concept of fellow feeling, or affective attunement, that the feelings of the participating individuals are numerically different facts. Talk of "attunement" makes sense only if there are at least two *relata* that can be in (or out of) tune. There never was (and never will be) *one token feeling* with many parts and participants in these cases, but only individual feelings, perhaps with mutual cognitive and affective attitudes. This is what makes talk of sharing a feeling merely *metaphorical* when applied to these cases.

The question I wish to pursue in this chapter is this: can feelings be shared in something like the straightforward sense of the word?

Immediately following the passage I quoted above, Scheler suggests that we can. Let me repeat the quote:

> Father and mother are standing at the dead body of a beloved child. They feel 'the same' grief, 'the same' pain. This does not mean: A experiences this grief and B experiences it, too, and in addition to that they know that they feel it – rather, their feeling is a feeling-together (*Mit-einanderfühlen*). A's grief is not in any sense an object of B's beliefs or feelings, as it is in the case of C, who approaches the parents with sympathy for them or for their pain. Rather, they feel their grief *together* in the sense of a feeling-together, an experiencing-together (*Miteinander-erleben*) not simply of the *same* value-state (*Wertsachverhalt*), but of one and the same emotional impulse (*Gefühlsregung*). [A's grief and B's grief are] *not two different facts*. (Scheler [1913] 1974: 23–24; my emphasis)

Here, Scheler does indeed claim that, in this case, the feeling of grief is shared in something of the bluntest straightforward sense of the word: *one* (token) feeling, many participants. The claim clearly is not that these parents simply experience feelings of the same *type*, or "matching" individual feelings, perhaps together with some form of (mutual) knowledge, (mutual) sympathy, or fellow feeling. And neither is there any proper *sympathy* involved in this case. Rather, the claim is that, while both individuals experience a feeling of grief, there are not two feelings involved in this case, but only *one*. The parents' feeling of grief is *numerically identical*.

Needless to say, the idea that feelings can be shared in the straightforward sense of the word is quite unusual and provocative. Indeed it might even seem that Scheler's case is simply too weak to be pursued. I take it, however, that in philosophy, unusual ideas should be highly valued and examined, even if their prospects for turning out to be true might appear rather dim. For, even if they turn out to be outright wrong at the end, such ideas might serve the purpose of helping us to clarify the *reasons* we have for rejecting them, and thus shed some light on why we conceive of matters in the way we normally do. This is what I shall be doing next: I shall use Scheler's claim that feelings can be shared in the straightforward sense to examine the reasons we might come up with in order to defend our strong pre-theoretic (or folk psychological) intuition that there is something wrong with this idea.

I shall use the label "individualism about feelings" for the view that feelings cannot be shared in the straightforward sense of the word. I take it that individualism about feelings is a deep-seated conviction, that permeates not only most theories of affectivity, but large parts of our pre-theoretical views (i.e. folk psychology) as well.

At least three different versions of individualism about feelings can be distinguished. First, feelings are *ontologically individual*. (Feelings are *conscious* states. As such, they are *ontologically subjective*, i.e. *somebody's* feelings, and there seem to be no conscious subjects other than individual beings.) Second, feelings are *epistemically individual*. (If it is true that *only individuals* can have feelings, it is also true that individuals can have *only their own* feelings.) Third, feelings are, it seems, *physically individual*. (If it is true that individuals can have only their own feelings, it also seems to be true that individuals experience their feelings *as localized in their own bodies*. This seems to force us to individualize feelings in the exact same sense as bodies; under normal circumstances, however, the only bodies there seem to be are the individuals'.)

Let us now have a closer look at each of these versions of individualism about feelings, starting with the last version.

Up until now, we have not discussed the nature of feelings beyond stating that they have a qualitative, phenomenal nature, i.e. that there is something "it is like" to have them. So let us have a closer look at what kind of phenomena might count as feelings. In his *Concept of Mind*, Gilbert Ryle gives a list that is often quoted: "thrills, twinges, pangs, throbs, wrenches, itches, prickings, chills, glows, loads, qualms, hankerings, curdlings, sinkings, tensions, gnawings, and shocks" (Ryle 1949: 83–84). This list is particularly well suited to emphasize the individuality of feelings. Clearly, there has to be somebody who *experiences* these phenomena – a role for which individuals seem to be the only suitable candidates. And clearly, the itches, throbs etc. individuals feel are *their own* – they are *epistemically exclusive*, as it were. Moreover, individuals experience the feelings they have *as localized in their own phenomenal bodies*. Feelings are body-related. A twinge is felt as a twinge *in the stomach*, the "throbbing feeling" is *in the breastbone*, and so on. Feelings are felt as localized in the body. In that, feelings are spatiotemporal entities. This is also illustrated by the fact that feelings can be, to some extent at least, co-present. A twinge in the stomach is compatible with a throbbing feeling in the breastbone, and similarly for other cases.

How does localization in the body amount to individuality? In the normal case, phenomenal bodies are *individual* in a twofold sense. They are individual *in themselves* (as unified fields of experience), and they are individual in that they are related to our physical bodies, and normally, physical bodies are individual. In themselves, phenomenal bodies are individualized by their *mode of access*: nobody feels a twinge in somebody else's stomach, let alone a throbbing feeling in some collective *Leviathan*'s breastbone. That is to say, phenomenal bodies are *our own* bodies, and they are individual, not collective. Phenomenal bodies are individual in their relation to physical bodies, because in the standard case (i.e. if we disregard for the moment phenomena such as phantom pains, which are parasitic in that their very possibility depends on the normal case), phenomenal bodies are co-extensive with physical bodies. And, if we disregard the extremely rare case of conjoined twins, and the possibilities of collective bodies such as *Leviathans*, physical bodies are the bodies of single individuals.

Let us call this form of individualism about feelings B-individualism. Tying feelings to our bodies, B-individualism seems to accommodate a deep-seated intuition, which might explain some of the reluctance in recent collective intentionality analysis against affective intentionality. Cognitions and pro-attitudes might be collective; feelings, by contrast, seem to be individual, because they are localized in individual bodies. Yet there seems to be one exception. Not only has Margaret Gilbert devoted more attention to the affective than all other participants of the current debate, taken together; she has also provided a standard formula for joint intentionality according to which the participants are "jointly committed *as a body*" (her italics). She thereby seems to suggest that intentionality has to be *embodied* in some way, and that shared intentionality therefore requires some kind of *shared body*. But in personal communication, she has made sufficiently clear that any such reading completely misses her point. Gilbert's *plural subjects* do not have bodies of their own. The formula of "being jointly committed *as a body*" should be read as a *metaphor*, which uses the unity of the individual's own body as an analogy for the collective unity that is generated through joint commitment. The collective "command center" of the plural subject (Gilbert 2000: 5) does not have any arms and legs other than those of the participating individuals.

For all of its apparent plausibility, there is something dubious about B-Individualism. The problem is not that there are collective phenomenal or physical bodies. Rather, the problem is that not all feelings are body-related in the sense of being localized in the phenomenal body. There is an alternative to this individualizing reading of feelings. In the course of the current *phenomenological turn* in the philosophy of emotions, the focus of attention has shifted from those feelings, which are bodily sensations, and has turned to what are usually called *psychic feelings* (Stocker 2003). In contrast to bodily sensations, these feelings are much more a matter of the *soul* than of the body. Very much along these lines, Descartes says of some passions that they are felt "as if they were in the soul itself". A rather useful distinction between bodily feelings and psychic feelings can be found in Scheler ([1916] 2000: 335–345). It is telling that Scheler limits his concept of "immediate feeling-together" (*unmittelbares Miteinanderfühlen*) to *psychic*

feelings. Following Scheler, psychic feelings are, in contrast to bodily feelings, intentional and not localized in the body, even though they are not without any relation to the body. The peculiar "heaviness" of the feeling of grief – the paradigmatic case of a psychic feeling in Scheler – does have a physical aspect. But grief is not *localized* in the body, as if one's grief were felt in one's left leg. According to Scheler, the peculiar nature of psychic feelings is such that they do not constrain the possibility of such feelings' being shared in any way. Remember that Scheler's example for sharedness of feelings in the straightforward sense is the parents' grief for their deceased child. Scheler suggests that this feeling has to be individualized in a way different from the participating bodies: *one* (token) feeling, two bodies. As opposed to what some of Scheler's interpreters claim (van Hooft 1994), this identity is not just a matter of the *content* of the feeling, but a matter of the identity of the feeling *as an emotional impulse* (*Gefühlsregung*).

Even if we grant that there are indeed such things as psychic feelings, and that the relation of this type of feelings to the body is not of a kind that leads to B-individualism, there are at least two further reasons for doubting that sharing of feelings in the straightforward sense could be possible. Let us first turn to what one might call *ontological individualism about feelings* (or O-individualism). The claim is the following: as mentioned above, feelings are conceptually tied to consciousness. (This does, of course, not mean that they are either *permanent*, or the focus of attention, or that one cannot be mistaken about one's feelings. None of these claims is implied in the definitional consciousness of feelings.) Consciousness, however, is *ontologically subjective*, i.e. it is *somebody's* consciousness. There has to be someone who *has* it – the good old subject. It seems, however, that individuals are the only members of the class of subjects of consciousness we know of. Therefore it appears that only individuals – and not collectives – can have feelings.

Ontological subjectivity does not only concern conscious states. It is a mark of the mental as such. Even those intentional mental states which are not conscious – such as non-occurrent beliefs and intentions – are ontologically subjective. Thus it is not surprising that the question of the subjectivity of collective intentions and shared beliefs has already stirred up considerable debate in the received literature on collective intentionality. The point at issue is the following. If it is claimed that there is such a thing as collective intentionality, and if it is further claimed that collective intentionality is in some way *different* from individual intentionality, we are pressed hard to think that, while individual intentionality is "had" by individuals, the subject of collective intentional states are collectives. Yet there is considerable resistance to this apparently innocent assumption in current collective intentionality analysis. Most authors explicitly refuse to ascribe any mental states to collectives. Shying away from the group mind assumption, authors such as Michael E. Bratman and John R. Searle, resort to one or another version of a *distributive* reading of collective intentionality. Bratman claims that collective intentions are basically a web of individual intentions with collective content, while Searle, for his part, claims that collective intentions are irreducibly collective in mode, but "inside" the mind of individuals. In order to ban the group mind, some form of *individualism* about intentionality is adopted (cf. Chapter 2 above). The central insight behind this seems

to be the worry that, if any subject other than the participating individuals were the subject of the collective intention, the subjectivity and agency of the individuals involved in the process would be somehow impaired or compromised.

It is true that individuals are and remain intentional agents when they join their forces for the purpose of collective actions, and they remain individual 'cognizers' when they share a belief. Individuals do not act and hold beliefs "on remote control", as it were, when they share a belief and act jointly. Participants of collective intentional states are *agents* and *cognizers* rather than mere parts, organs, or instruments of some unified collective mind. Each individual participant can be ascribed beliefs and desires, which rationalize his or her behavior. In other words: any conception of collective intentionality should be compatible with the *intentional autonomy* of the participating individuals.[9]

Although individual intentional autonomy might *explain* why most philosophers of collective intentionality shy away from the group mind, it is by no means a *good reason* for doing so. There are robust conceptions of the group mind, which are perfectly compatible with individual intentional autonomy. This is particularly obvious in Philip Pettit's work. In his analysis of the *discursive dilemma*, Pettit has developed the view that (certain types of) collectives can be ascribed some sorts of mental states and do belong to the class of intentional subjects. He even ascribes them some sort of personhood (Pettit 2002; cf. Rovane 1998). In this sense, groups can have "a mind of their own" (Pettit 2003). In his analyses, Pettit's concern is with the *rational unity* of the groups' perspective, which under some circumstances require a peculiar *discontinuity* with the participating individuals' own perspective. The phenomenon to which Pettit draws our attention is that the rational unity of a group perspective sometimes requires that this perspective be *distinct* from that of any of the participating individuals'. At the same time, Pettit argues that the group mind is solidly grounded in the volitions of the participating individuals. Groups have minds of their own based on the participating individuals' *insight* into the problems of aggregating individual decisions to collective decisions, and on the participating individuals' *desire* to act consistently and rationally, as a group. It is telling that while Pettit endorses the group *mind*, he stoutly rejects any notion of group *consciousness* (Pettit 2002: 443). In his view, collectives might be ascribed some collective subjecthood in that they have mental states of their own; collectives might be treated

[9] It is important, however, to distinguish the assumption of individual intentional autonomy from two further claims, with which it is usually mixed up. Individual intentional autonomy is the claim that, in normal cases, each individuals' behavior instantiates *his or her own action* (this excludes *intentional heteronomy*, i.e. behavior "on remote control"). The assumption of *individual motivational autarky*, by contrast, states that *each individual ultimately acts only on his or her own desires*, i.e. the intentional explanation of each individuals' behavior should bottom out in that individuals' own pro-attitudes (this excludes *motivational heterarky*, i.e. the possibility of acting on other people's pro-attitudes without making them one's own, or acting on another corresponding pro-attitude of one's own). The third claim, intentional individualism, is weaker than motivational autarky. Intentional individualism is the claim that people can act only on *individual* intentions (this excludes *intentional commonality*, i.e. the existence of intentional states which are shared in the straightforward sense). In Schmid 2007, I claim that individual intentional autonomy does not imply motivational autarky and intentional individualism (cf. Chapter 1 above).

as *responsible* for their actions, and thus be ascribed some sort of personhood. But there is no "what it is like" to have these states and to be these collective persons. Collective minds and persons do not involve any qualitative, phenomenal dimension, because collectives do not have any *consciousness* of their own. The consciousness of the participating individuals is the only consciousness there is. Thus even those philosophers who are liberal with regard to the collective mind exclude the possibility of collective feelings; this is the immediate consequence of their ruling out the possibility of collectives exhibiting some sort of consciousness in the first place.

Let's now turn to the third and last type of individualism about feelings. It is *epistemological individualism* (or E-individualism). If it is true that only individuals can have feelings, it is also true that the feelings individuals have are *their own*. Feelings are not only *subjective*. They are *epistemologically exclusive*, too. This is not to say that individuals are infallible as to the content and kind of feelings they have. In fact, we often have reason to revise our interpretations of our own feelings in light of other people's interpretations. But there is a way in which my feelings are accessible to *me* but not to anybody else. The difference in question is usually called the difference between inferential and non-inferential knowledge. By contrast to outside observers, I do not have to read any signs in order to know what I feel – I just *feel* it. And even though I still might misinterpret my feelings, it seems that the particular way in which my feelings are accessible to me gives me a special kind of *authority* concerning the nature of my feelings.

This asymmetry in the way one's own and other people's feelings are accessible to us seems to preclude the possibility of shared feelings in the straightforward sense. If E-individualism about feelings is true, shared feelings cannot consist in a numerically (or token) identical feeling, for if people were to share a feeling in the straightforward sense, it seems that they would have to *have each other's (token) feeling* (and not just a *similar* or *suitably related* feeling of their own), which is not compatible with the epistemological exclusivity of feelings.

Again: it is clear that individuals can have similar or perhaps even qualitatively identical feelings. And it is undisputed that other people, or other people's feelings, can be in the *intentional content* of individual's feelings. E-individualism is not in conflict with any of the metaphorical forms of the sharedness of feelings. For those forms of affective attunement, sympathy, and empathy do not compromise the epistemological exclusivity which E-individualism is all about. Quite the contrary: those forms of sympathy and fellow feeling *presuppose* that people have only *their own* feelings and that one person's feeling is numerically distinct from another person's. This is particularly obvious in simulation theories of empathy and sympathy, and in Robert Sugden's view of Adam Smith's concept of *fellow feeling* (Sugden 2002). As mentioned above: if *fellow feeling* is a matter of people enjoying their mutual affective *attunement*, it is clear that the feelings have to be numerically distinct, because it takes at least two feelings for there to be any attunement (or indeed disharmony) of feelings. There is no "fusion" between the feelings of the participants in these metaphorical cases. The feelings of the participants are (and remain) "different facts", which is not the case in the straightforward sense of sharing a feeling.

E-individualism about feelings enunciates our deep-seated conviction that we cannot know how other people's joys and pains *really feel*, i.e. *what it is like* for other people to feel pain, or to experience joy. Our notions of sympathy, compassion and empathy – just as our philosophical theories of empathy from Theodor Lipps up to current simulation theory – are marked by the insight that those feelings and beliefs *never really reach what they aim to grasp*. There is no better illustration for our convictions concerning the *inaccessibility of each other's feelings* than our recognition of other people as the ultimate epistemic authority concerning their own feelings. But still, there are some divergent views to be found in everyday life, literature, and philosophy. There is a way of talking about the relation between individuals that seems to imply that the monads do indeed have windows. It is perhaps no coincidence that Scheler chooses the parents' feelings towards their child as an example. These bonds might be so tight that, at times, it becomes difficult to distinguish one's own feelings from an other's. A literary illustration is provided in Nadine Gordimer's *The House Gun*, where the parents of an accused murderer constantly swap their rapidly changing feelings, so that it does indeed seem that some of their feelings are, in quite a literal sense, *not their own*.[10]

But the intuition concerning the possibility of sharing a feeling in the straightforward sense goes far beyond the range of intimate relationships. Taken with a grain of salt, one might say that even a presidential campaign has been won based on this claim. Remember that Bill Clinton cast an important part of his public persona with his claim not only to *know* what others feel, but to actually *feel it*. One might suspect Clinton's famous "I feel your pain" (which was so often repeated and endlessly quoted) of Frankfurtian *bullshit*. But even if this should be true, the question is: why and how does it *work?* It seems that Clinton's claim is at odds with E-Individualism. For it seems clear from the context that Clinton's statement cannot be understood in some of the metaphorical senses of sharing a feeling. Clinton does *not* mean that he *understands* the other person's pain ("She doesn't *feel* your pain, she *understands* it" – according to the influential political commentator Joe Klein, that's precisely the difference between Bill and Hillary). Does Clinton mean that he *knows* about the other's pain, and that he feels sorry about it, is saddened by it, or that he has some other sympathetic feeling of sorts? This is rather implausible because of the wider context in which Clinton made his statement. The place was a Night Club in Manhattan, the time was the evening of March 26, 1992. With his statement, Clinton addressed Bob Rafsky, an Aids activist who had disrupted Clinton's stump speech rather violently by voicing his anger at the government's apparent inactivity in that domain. With his reply "I feel your pain", Clinton does not mean that he *knows* about Rafsky's pain (which would amount to no more than stating the obvious), and that he is saddened by the fact that Rafsky is in pain. Clinton's concern is clearly not so much Rafsky as the disease. It is the suffering caused by the disease and the fact that the government neglected the issue that pains him. Does Clinton therefore simply mean that Rafsky has no monopole of being pained by the disease, and that he, Clinton, feels similarly or even identically when it comes to aids? It is true that,

[10] I am grateful to Margaret Gilbert for drawing my attention to Gordimer's novel.

according to the transcripts of the incident, Clinton mentions that close friends of his had died of aids, thereby suggesting that aids is the source of a great deal of personal pain to himself. But still, this reading seems implausible, because Rafsky had made clear that he himself is terminally ill of aids. So how could Clinton ever compare his own pain to Rafsky's? Bullshit or not: what could Clinton's statement possibly mean?

We should not dismiss the possibility of the plain and simple straightforward reading: Clinton means that he feels Rafsky's pain; he does not mean that he sympathizes with Rafsky, or that he has a similar feeling of pain of his own, but simply what he says: that he feels Rafsky's pain. This would be an asymmetric case of sharedness in the plain and straightforward sense: the pain is – and remains – primarily Rafsky's; but Clinton *takes part* in Rafsky's pain in something of the sense in which Clinton entered the room in which Rafsky was waiting when he came to deliver his stump speech: one Night Club, two guests, one state of pain, two subjects. This reading seems appealing enough – only that it is incompatible with E-individualism about feelings.

To sum up this discussion of the three versions of individualism about feelings, it seems that, while there are some rather strong intuitions concerning the individuality of feelings that should not be dismissed light-heartedly, it might still be worthwhile to consider alternatives. Let us therefore have a second look at the notion of straightforward sharedness as it seems to be implied in Scheler's view. Scheler seems to reject P-, O- and E-individualism alike. As for E-individualism, Scheler advances his *perception theory of the consciousness of others*, in which he claims that there is no fundamental difference between our access to our own consciousness and to that of others. There is, Scheler claims, an "inner perception" of the feelings of others that leads us to experience other people's feelings in just the way in which we experience our own feelings (Scheler 1974: 283), which is what E-individualism denies.

Scheler's further comments on the matter lead him into conflict with O-individualism, too. While he does not claim that there are any collective subjects of consciousness, he simply denies the very idea of the ontological subjectivity of consciousness as such, stating that, in its original form, the stream of consciousness is "quasi anonymous" in character. The distinction between my and your consciousness is not fundamental, as Scheler suggests, but only the result of a process of differentiation. So at the basic level of consciousness, it is not anybody's, and therefore cannot be ontologically subjective.

This, however, seems to be in blatant conflict with a widely shared intuition that is at least as deep-seated as any of the other intuitions mentioned above. It is the idea of the *separateness of individual persons*. Contrary to what some philosophers (such as Arthur Schopenhauer; cf. [1841] 1988) thought, the fact that we are separate persons is not only a matter of *appearances*. Our pre-theoretic intuitions, our theories, and much of literature are based on the assumption that we really *are* separate persons, that the otherness of other people is not just a superficial "construction". This intuition, too, should not be dropped light-heartedly. Love lyrics – just to quote one example – nicely capture our ambivalence concerning jointness and separateness between persons. Here, the idea of merging, of fusion, is pervasive; at the same time, it

is full of vivid descriptions of how the innermost of the feelings and experiences of other persons elude us, how some permanent *fusion of feeling* remains unattainable however hard we might try. Thus the question is: is there any way to reconcile these two intuitions, i.e. the straightforward concept of sharedness of feelings on the one hand, and the intuition concerning the separateness of persons on the other?

§15 Phenomenological Fusion

In Scheler's remarks, as well as in the short but lively debate that was ignited by Scheler's proposal, a solution to the dilemma seems to emerge. The point of departure is a structure that does not seem to have received sufficient attention on the part of analytical philosophy as yet. It is the fact that conscious states are not just ontologically subjective, in the sense that they are to be attributed to someone who "has" it. To be in a conscious state also means to have some pre-reflective *awareness* of one's being in a conscious state. In however low a degree of clarity, to be in a state of consciousness involves some self-referential awareness. Two clarifications are in order. First, this does not mean that one cannot be wrong about one's consciousness. It seems obvious that we are very prone indeed to misinterpretations of our own conscious states. Second, the claim that to have a conscious state implies some awareness of one's being conscious does not mean that in order to be conscious, subjects have to reflect on themselves all the time. Self-reflection only serves to make explicit the peculiar pre-reflective awareness characteristic of any kind of consciousness, and is very different from that awareness.

Conscious states are – pre-reflectively and un-thematically – *conceived* and *interpreted* by the subjects who have them. To be in a conscious state means to *conceive of this state* in some or another way. In the case of intentional states of consciousness, this concerns the intentional *content* as well as the *mode* and the *subject* of the intention. In the latter respect, to be in an intentional state of consciousness means to conceive of oneself pre-theoretically and un-thematically as the subject of that intentional state. Consciousness conceptually involves some form of *self-awareness*, some conception of oneself as the one who "has" the consciousness in question. This brings in an important distinction. The subject of a conscious state can mean either of the following: (a) the subject who has the conscious state in question; (b) the subject *as who* the subject takes himself or herself to have the state in question. If we allow for the possibility that (a) differs from (b), this seems to open up a perspective in which to make the two conflicting intuitions compatible. Of course, individuals can have only their own conscious states, especially feelings – but this does not answer the question *as who* those individuals take themselves to have their conscious states. In other words: O- and E-individualism might be true with regard to subject_A, but be wrong with regard to subject_B of the consciousness in question. And our intuition concerning the connectedness of persons might be true with regard to subject_B, but not with regard to subject_A. Without doubt, the parents in Scheler's example are two different persons each of whom has his or her own feelings. But

this does not preclude the possibility that both of them experience their feelings as theirs (together) rather than as separate personal feelings.

The following phenomenological consideration seems to lend further support to this assumption. It seems that in everyday life, we experience only very few of our conscious states *as our personal conscious states*. In fact, it seems that we take our conscious states to be our own only where we have reason to think that our conscious states might be different from anyone else's. Where this is not the case, we simply think what *one thinks* or *what is generally thought*, in an a personal or anonymous mode, as it were. We do not take our thoughts or feelings to be *our own* in any meaningful sense. In other cases, we take our conscious states to be those of *other people's*, such as in cases of Lippsian "internal co-action", in which we identify with other people. In these cases, the "metaphysical psychic substance", as Scheler puts it – i.e. subject_A – is individual. But subject_B differs from subject_A.

This is the direction in which Scheler's remarks – together with the (admittedly very few) defenders of Scheler's view – seem to point. The question is: how much weight can be put on the structure "S_A feels x *as had by* S_B"? Does it really make good sense to allow for a systematic difference between S_A and S_B in order to make room for cases such as that of strong sharedness? It has the disadvantage of hinging on a poorly analyzed structure, and it raises the following criticism: are cases in which the two levels of subjectivity diverge, not simply cases of *self-misinterpretation*? Why not say that people are simply *mistaken* whenever they take their feelings to be anybody else's rather than their own personal feelings? Isn't it obvious that people who take their feelings to be some other individual's, or some collective's, or nobody's in particular, are on the wrong track? And this is only the first of a whole series of critical questions.

I think that a valid, straightforward notion of sharedness should conform to the following four conditions of adequacy (the conditions are drawn from the debate over Scheler's account in the German speaking world of the late 1910s and 1920s).[11]

1. *The self-awareness condition*. Any notion of sharedness must be compatible with veridical self-awareness. It cannot require that one should *mistake* oneself for another person, or for no person at all, or for a collective.
2. *The fallibility condition*. It must be compatible with the insight that one can always be mistaken about other people's feelings. No feeling, however strongly felt, and however intimately connected one believes it to be to another person's life, provides infallible information about other people's feelings. In other words: no feeling is in itself the criterion of its being shared. Even parents might be mistaken in their belief that their feelings for their child are shared. (A particularly touching example from the *belles letters* can be found in Thomas Hürlimann's *Das Gartenhaus*. Here, again, we are introduced to a couple having lost their child. The child's mother reacts with shock when she finds out that, contrary to

[11] The participants include Erich Becher, (1917, 1921), Ludwig Binswanger (1922), Karl Bühler ([1927] 2000), Johannes Cohn (1919), G. Roffenstein (1926), Max Scheler ([1912] 1954, [1912–16] 1973), Alfred Schütz ([1931] 1991), Edith Stein (1917) and Johannes Volkelt (1920).

what she thought, her grief is not their *shared* grief after all, and that for years the father's sole reason to accompany her to the child's grave has been to feed the graveyard cats.)[12]

3. *The holism condition*. Feelings, as with any intentional conscious states, do not occur in isolation. Rather, they depend on each other within the total stream of consciousness (Edmund Husserl calls this the "horizontal structure" of intentionality). So an adequate account of sharedness has to be able to explain how, in spite of the holistic structure of consciousness, it is possible to share some intentional states, without sharing the total stream of consciousness.
4. *The difference condition*. In many cases, sharing a feeling is a matter of the differences between the feelings of the participating individuals rather than any form of identity. If Clinton claims to feel Bob Rafsky's feeling, he certainly does not mean that he, Clinton, a presidential candidate, feels Rafsky's pain – a terminally ill aids-activist's – in the *same intensity*, and indeed in the same *quality* as Rafsky. By way of another illustration, take the following example: imagine the shared feeling of joy at the success of the first performance of a symphony. If the man at the triangle, the composer, some member of the audience and the stage manager take themselves to share a single feeling of joy, this is because, in their perception of the situation, their individual feelings "match" with that of the others rather than being qualitatively or even numerically identical. In order to be taken as "matching", these feelings have to be taken to be *different from each other* according to the different roles the participants play in the joint activity. If the composer takes the man at the triangle and the member of the audience to share her joy, she will not, in her right mind, take them to experience her exuberant exaltation; rather, she will take the shared feeling to entail her own exuberant elation together with, for example, the audience member's delight, and the man at the triangle's silent satisfaction. Thus taking part in a shared affective state sometimes seems to entail some form of awareness of the *difference* between the feelings of the participating individuals rather than any awareness of identity.

At least, in their general thrust, all of these points seem convincing enough. It is, I believe, reasonable to say that any concept of sharedness of feelings, however straightforward, should be able to meet the following requirements. Firstly, it has to be compatible with basic forms of individual self-awareness. People do not have to mistake themselves for another, or feel completely dissolved in some group consciousness in order to share a feeling. Secondly, it has to be compatible with the knowledge that any feeling one takes to be shared might not actually be shared at all. Thirdly, it has to leave room for the experience of (partial) separateness of our conscious lives. Not all feelings are shared. And ultimately, it has to conform to the experience that very often (if not always), the sharedness of a feeling is a matter of the qualitative difference between the individual contributions.

Can any straightforward conception of shared feelings possibly meet these conditions? I think it can be shown that it can. Let me start with the third point. The

[12] I am grateful to Angelika Krebs for pointing me to Hürlimann's novel.

main problem the holism condition seems to be driving at is the following. If conscious states depend on each other within a horizontal structure, and if one role of the concept of the *subject of consciousness* is precisely to capture this holistic form, consciousness can be *either* collective *or* individual, but cannot be switched *ad libitum*. If *one* conscious state were to be straightforwardly collective, *all* would have to be collective.[13] I do not think, however, that this argument poses any serious threat to a straightforward notion of sharing. Even if we grant that the general thrust of the argument is right, it is overly restrictive since there are not just single conscious states on the one hand, and the total stream of conscious on the other. There are intermediate structures: *episodes* of our conscious lives. While it is true that the experience of one single feeling will not be taken as shared in complete isolation, this does not mean that one's total stream of consciousness would have to be shared in order for one element of an episode to be shared. It is within shared intentional *episodes* that these phenomenological fusions of feelings occur, and presumably, such episodes require some form of common life and shared practices. One does not take oneself to share feelings with completely strange creatures. Strong sharedness requires a context, perhaps some form of *intimacy* between the participants, or at least some *shared cognitive and conative attitudes*. Episodes may differentiate shared conscious states from other conscious states that are had in separation from others. In this way, the straightforward sense of sharedness seems to be in tune with the fact that we do not share our entire conscious lives. The fusion – if there is any – is only partial.

This brings us to the second point in the above list. I expect that an analysis of the structure and presuppositions of these episodes should also yield some insights into the independent truth conditions to which a feeling has to conform in order to count as shared in the straightforward sense. This is an issue not to be pursued further here. It is clear, however, that whatever we might feel when we share a feeling cannot itself be the criterion of the truth of the sharedness, and that we can always be wrong about our feelings actually being shared. This does not preclude, however, strong sharedness in cases where these conditions (whatever they might be) are met. Furthermore, it seems clear that one does not have to mistake oneself for another person (let alone some group consciousness) to experience her or his feelings, if one's feelings are taken to be *shared* with that person. In other words, the individuals experiencing a feeling as shared can be aware of the difference between what we have called subject_A and subject_B above. Thus it seems that the self-awareness condition can be met, too.

The most serious attack on the idea of a phenomenological "fusion of feelings", however, comes from the last of the abovementioned arguments. Remember that the problem is this: if we grant that, in most if not all cases of shared feelings, the experiences of the participating individuals are taken to be qualitatively different by the participants qua "matching" contributory feelings to a shared collective feeling, it seems plausible to see these feelings as numerically distinct *qua* suitably matching

[13] If I understand him correctly, this is the upshot of Alfred Schütz's critique of Scheler's views (Schütz [1931] 1991: 147).

parts to a whole. It is essential to the shared feeling of joy at the success of a performance that the exuberant exaltation of the composer is *not* the delight of the member of the audience, or the silent contentment of the man at the triangle, etc. This, however, seems to necessitate the abandonment of any straightforward notion, in two ways at least. Firstly, if the feelings of the participating individuals are conceived of as (matching) parts, they are *different from each other*, not numerically identical (just as a bicycle's frame cannot be its wheels if it is to serve its function within the total structure of the bicycle). And, secondly, if the shared feeling is taken to be a whole with parts, it seems to be different from the parts – in other words, a collective "we" beyond the individual "I" and the "thou" comes into play as the proper subject of the shared feeling.[14]

I think, however, that these metaphysical conclusions can be avoided (in part, my solution is inspired by composition-as-identity theory as developed by Donald Baxter, David Lewis and David Armstrong). Firstly, if self-awareness is indeed compatible with straightforward sharedness, it seems that there are *two ways of counting* the number of feelings involved. With regard to subject_B (which is a "we"), the number is *one*. With regard to subject_A, the number is *two* (in the dyadic case). *There is no reason why one way of counting should be more legitimate than the other*. At the same time, however, there is no legitimate way of counting that yields *three*. The feelings can be counted *either* by phenomenological subject, *or* by ontological subject, but to count the "we" as a *third subject on the list* seems about as futile as the attempt to make some extra money by trying to sell one's coin collection *as a whole* after one has already sold all its parts, or to wait for the teams to appear on the field after all the players have entered. The shared feeling is nothing *in addition* to what the participating individuals feel. Rather, it *is* that feeling, and it *is* that feeling *in a certain respect*. The individuals' feelings *are* the one shared feeling insofar as the conditions under which individuals are not mistaken in experiencing their feelings as being shared by the other participants are met.[15] This leads to a robust notion of a *collective consciousness* that avoids the pitfalls of some forms of anti-individualism in that it does not force us to reject either O- or E-individualism.

The first of the challenges posed by the argument from difference, however, is still not met. It is that sharedness in the straightforward sense seems to be incompatible with the awareness of differences between the feelings of the participating individuals. It seems that, in most cases at least, the participating individuals have to take their feelings to be *different* from the others' feelings when they take themselves to share an emotional state. I do not think, however, that this disproves the claim that

[14] These are indeed the two main revisions demanded by Edith Stein (1917: 18) in her discussion of Scheler's view.

[15] It should be at least noted in passing that the main problem of this view (which conforms to the composition-as-identity theory as put forward by David Lewis, David Armstrong, and Donald Baxter; cf. Baxter 2005) is that, whereas the shared feeling is *one*, the individual feelings are *many:* "what's true of the many is not exactly what's true of the one" (Lewis). Composition-as-identity-theory might be controversial. But the mere fact that it is around should remind us not to give up a corresponding pre-theoretical phenomenological intuition too early just because it does not seem to fit easily with our metaphysical views.

people can consistently take their feeling to be token-identical with the others when they share a feeling. Sharing in the straightforward sense leaves ample room for difference qua *aspects* of the whole: the shared feeling of joy after the successful first performance of a symphony, *insofar as* it is felt by the composer, is one of wild exuberance; *insofar as* it is the man at the triangle in the orchestra's, it is rather one of silent contentment – still, it is *one* feeling. Thus the numerical identity of the feeling does not preclude difference, but the difference in question here is one between *aspects* of one feeling rather than one between numerically different feelings.

In a paper from 1997 – 5 years before she put forth the views on the structure of collective emotions that I mentioned above – Margaret Gilbert pointed out the following phenomenon. Individuals can feel guilt or pride for the actions of their groups, even if they did not serve any part whatsoever in the respective group activity. Their feelings, Gilbert claims, are of a different quality than the feelings of guilt or pride individuals may feel regarding their own participation in joint actions. Gilbert has labeled the former, non-personal type of feelings *membership feelings*. Membership feelings, Gilbert claims, are an important feature of collective emotional states.

In a paper from 2002, however, Gilbert moves away from this earlier, largely phenomenologically inspired account. She quotes two reasons for her dissatisfaction with her own earlier views. First, *membership feelings* lack the unity of genuinely collective emotional states; they have no collective subject in the sense of a subject that can be distinguished from the participating individuals. Membership feelings appear to be purely individualistic, i.e. not of the same cloth as plural subjects. Secondly, and perhaps more seriously, Gilbert points out that membership feelings are not sufficient as conditions for ascribing emotional states to collectives. If there is no common knowledge of the respective feeling among the participants, it is possible that *all* individuals feel membership guilt without there being any shared feeling (Gilbert 2002: 135ff.). In this case, the collective cannot be ascribed any guilt feeling, even though all members feel membership guilt. These seem to be the two reasons why Gilbert takes a cognitivist line on the matter in her more recent thought. In the course of her development, she turns away from phenomenology. Gilbert now thinks that the phenomenal aspects are no more than contingent accompaniments of collective emotional states, and that an account of collective emotions should be based on an analysis of the shared cognitive and practical components of these emotions, rather than on a phenomenology of the feelings.

I hope that, over the course of this chapter, it has become clear that neither of the two reasons Gilbert quotes for her cognitivist turn holds water. The problem of her earlier account is not that it is based on feelings, but that the feelings are conceived of in an overly individualistic fashion. A more straightforward notion of feelings avoids the problems Gilbert takes as the reasons for her turning away from phenomenology. First, if feelings are shared in the straightforward sense, they do have a collective subject, which is different from the subject of the participating individuals (i.e. subject_B). And, secondly, some structure of mutual openness is an integral part of sharing in the straightforward sense – remember Scheler's remarks on the parents' not needing any additional mutual awareness in addition to their

shared feelings, because the mutual openness is part and parcel of the very feeling itself (which does not mean, however, that the participants might not be mistaken). My proposal thus is to understand *membership feeling* as *shared feelings* in the straightforward sense. Gilbert's turning away from phenomenology is based on a mistake that is due to an individualistic understanding of feeling.

Feelings can indeed be shared in the simplest sense of the word. Our basic intuitions concerning the individuality of feelings leave ample space for a straightforward understanding of the sharedness of feelings. If this is true, phenomenology will be essential for an understanding of collective affective intentionality.

Part II
Collective Intentionality in the Social Sciences

Chapter 5
Social Identities in Experimental Economics

In a graphic image used in the introductory chapter to the *Foundations of Human Sociality*, Ernst Fehr and Colin F. Camerer compare the role and scope of experimental games in the study of human sociality to that of a first sketch or outline in the process of an artist's conception of a painting. Just like a rough draft, experimental games are, as Camerer and Fehr put it, "reductions of social phenomena to something extremely simple" (Camerer and Fehr 2004: 85). By abstracting from contingent details and by reducing complex phenomena to some of their basic (and perhaps essential) features, experimental games allow for "comparability across subject pools" (ibid.: 84), a feature of which the volume from which the quote is taken is itself a most impressive example.

Yet it has to be said that reduction is always dangerous, and the art of drawing offers an excellent example for the adventures of simplification: if too many details, redundancies and apparently contingent features are left out, people might misunderstand the sketch. In his Little Prince, Antoine de Saint-Exupéry has the narrator learning this the hard way when the little boy draws his 'Drawing Number One' – a giant snake digesting an elephant – only to learn to his disappointment that the adults mistake the lumpy blob with two lines tapering off to both sides to be a rendering of a hat! The lesson is that for all their simplicity, rough outlines need more interpretative work from the side of the beholder than more detailed pictures. And the more reductive a rendering, the more easily it is misunderstood. The question is: could this also be true of game experimental 'sketches'? How do we know what an experiment is about in terms of real-life social phenomena?

In the following, I shall discuss the most famous of Fehr et al.'s game experiments, his third party punishment-experiments which are designed to demonstrate how agents provide a second order public good by punishing unilateral deviation from the cooperation norm even where these agents themselves are not directly affected by the outcome. It will be argued that the interpretation of the resulting sketch given by Fehr et al. is misleading, and that the main problem of the "sketch" offered by Fehr's experiment is that it does not sufficiently account for the participant's social identities.

H.B. Schmid, *Plural Action: Essays in Philosophy and Social Science*, Contributions to Phenomenology 58,

§16 'Strong Reciprocity' and Other Misnomers

The structure of one version of the experiment at issue here (Fehr and Fischbacher 2004: 72ff.) is as follows. There are three subjects involved. Test person A is endowed with a certain amount of money by the experimenters, and she is offered the opportunity to spend parts of her fund on inflicting financial loss on another, randomly selected individual B (if A spends 1 unit, B's funds is reduced by, say, 3 units by the experimenters). A does not know who B is, she is assured anonymity by the experimenters, and she knows that there will be no further interaction whatsoever between herself and B in the future course of the experiment, so that A need not worry about her behavior influencing her reputation. Now this seems like a strange experiment indeed. Why should A be so keen on reducing B's funds the she is willing to spend some of her own money in order to achieve that goal, if B hasn't done anything to her beforehand?

There is something A knows about B which explains why she might choose that option. What A knows about B is B's decision in a previous round of the experiment, where B was in an interaction with yet another randomly selected individual C. In this first round of the experiment, both B and C were given the options either to keep the money they were endowed with by the experimenters for themselves, or to transfer their fund to the other, who would then receive three times the transferred sum (transfers are tripled by the experimenters). B and C had to make their decision simultaneously and without being given the opportunity to prior communication. Possible outcomes of the first round where the following: B ended up with 3 units if both B and C decided to transfer, he received 1 unit if neither decided to transfer, he was left with 0 units if he decided to transfer and C didn't, and he received the optimal amount of 4 units if he didn't transfer, but C did – and conversely for C.

After the second round (where A comes into play), the entire experiment is repeated several times over, with the test persons being regrouped for each new series, and each person being in the position of A in the second round playing the role of B in another group in the first round. No pairing, however, is repeated, so that the experiment consists of a series of one-shot interactions. The structure of the game is common knowledge among the participants.

Roughly stated, the results of the experiment are the following: A tends to spend a surprisingly substantial part of her endowment inflicting financial loss on B, especially if in the earlier round B did not transfer his money, while C transferred hers (A's tendency to inflict loss on B was low if either B decided to transfer, or if both B and C decided not to transfer). Anticipating A's negative reaction to unilateral non-transferral in the second round of the experiment, B was more prone to choose the 'transfer'-option in the first round than she or he was when the experiment was limited to the first round only (interaction between B and C with no intervention by A).

To use Fehr's and Camerer's metaphor, this is the outline. The decisive question now is: what does it show? What is the sketch about? What sense can and should be made of it in terms of real-life social phenomena? What is the deeper meaning of the behavioral pattern revealed by this experiment, and their place in social life?

Here is what Fehr et al. like to see in the sketch which their experiment produced. In Fehr et al.’s view, this sketch represents the provision of a second order public good. The label they use for A’s behavior is “altruistic punishment”, and they claim it to be ‘pro-social’ and ‘norm-enforcing’, and a paradigmatic case of ‘strong reciprocity’. But there is obvious reason to doubt this reading. Even the labels Fehr et al. use do not seem to fit the outline.

(a) The label “reciprocity”, as attached to the behavior at hand, seems rather strange, since neither of the pairings is repeated, and B is given no opportunity whatsoever to retaliate against A.

(b) The term “punishment” is questionable, too. For obvious reasons, A’s intervention cannot be interpreted simply as ‘punishment’ for B’s refusal to transfer her fund to C in the first round of the experiment, since B’s decision not to transfer was only ‘punished’ if C decided to transfer. In other words, B’s non-cooperative choice was ‘punished’ by A only if it constituted a case of unilateral deviation from the cooperation norm. It is obvious, however, that B had no control whatsoever over whether his non-cooperative decision constituted a case of unilateral non-cooperation, since this depended on C’s decision, which was not known to B, let alone under her or his control. This makes the label “punishment” highly questionable. In ordinary language, ‘punishment’ is an imposition of a penalty on somebody for some wrongdoing on his part. Whatever counts as right or wrong in a given context, the term implies some act on the part of the punished person for which she or he is punished. A’s ‘punishing’ behavior, however, was not triggered by B’s choice but rather by the outcome of the interaction between B and C. As B cannot know how his decision matches with C’s, and has only partial control of the situation, the term “punishment” seems misplaced here.

(c) The label “altruism”, as applied to A’s behavior, is no more self-evident than “reciprocity” and “punishment”. It is true that the behavior does not conform to the selfishness assumption of classical economic theory, according to which, because of the costs to A, no infliction of loss on B should have occurred. Thus, the behavior in question does not appear to be egoistic in a narrow sense of the word. Yet this does not, in itself, make it a case of altruism. Most experimental economists seem to tend to portray deviations from the principle of narrow self-interest as being of a ‘benign’ or ‘pro-social’ kind. As the true heirs of the enlightenment movement and its positive picture of human nature, they have often found humans to depart from their self-interest out of altruism, inequality-aversion, or general fair-mindedness (most obviously in the case of test persons voluntarily sharing their funds in the dictator game). What Fehr et al. have in mind when they call A’s behavior altruistic seems to be something very much along these lines. As they see it, “altruism” refers to the fact that A’s anticipated behavior made mutual transferal more frequent among B and C, very much to B’s and C’s mutual benefit (remember that the transferred sums were tripled by the experimenters, so that if both chose to transfer, both parties were better off). By making unilateral non-transfer (i.e. the attempt to cash in the other’s money while keeping one’s own) less attractive, A’s presence ultimately benefited both.

> While being altruistic with regard to the group consisting of B and C, however, it seems that from *both a more narrow as well as in a broader perspective, A's behavior could be called destructive or aggressive rather than altruistic*. Let us take each of these perspectives, in which A's behavior appears as anything but pro-social. In a more narrow perspective, it is hardly an act of altruism to inflict loss on another person; B is the *victim of A's aggression* rather than the beneficiary of A's altruism; and in a broader view, it is far from obvious why the increase of the total cost of the experiment that resulted from A's presence (everybody got more) should be considered an act of altruism! After all, *somebody* had to cover these costs, too (presumably the taxpayers, as the experiment, carried out at a state university, seems to have been government funded). A's propensity to punishment might have helped to maximize the participant's total earnings; with regard to the non-participants, this is hardly pro-social. Thus it seems obvious that labeling A's behavior as 'pro-social' and 'norm-enforcing' is as questionable as the label "second order public good", for she seems to have supported the appropriation of the experimental fund for the private benefit of B and C, which, in itself, is hardly a socially desirable outcome.

To this second line of argument one might object that the deliberative process of the participants is limited to the experimental situation, so that questions such as where the experimental funds came from and what the stakes of some wider public in the current situation are were not of concern to the participants. This is undoubtedly true, and it is indeed part of the whole idea of experimental games, so it is understandable that the experimenters do not seem to have thought about this point either; but as an essential part of how the participants 'frame' game experimental situations, it is in itself a remarkable fact that is in need of explanation.

My claim is not that the general thrust of Fehr et al.'s interpretation of their experiment is entirely mistaken; rather, the point is that whether the behavioral pattern discovered by Fehr et al. is rightly called "beneficial" or "pro-social" rather than "aggressive" or "destructive" is not self-evident from the mere lines of the experimental sketch, i.e. from the choices made (and the payoffs received). The problem encountered here is similar to that of the interpretation of Saint-Exup'ery's 'Drawing Number One': in order to decide what sense to make of the sketch (snake or hat?), we need to have the right kind of background understanding of the situation at hand.

As to the structure of this background, I put forward three interrelated claims:

1. The structure of the behavior in question cannot be adequately described within the conceptual distinction between behavioral egoism and altruism.
2. A strong concept of group-relatedness is needed in order to make sense of the observed behavior.
3. An adequate account requires us to depart from the atomistic notion of the agents' identity as implied in much of standard economic theory.

I shall proceed as follows. In the next paragraph, I approach the limitations of the distinction between egoism and altruism from a historical perspective, before introducing the role of social identities in the following paragraph (§18). In the

concluding paragraph §19, I develop a fuller account of Fehr et al.'s experiment – a sketch that detailed enough so as to make the social phenomenon at issue recognizable.

§17 Beyond Egoism and Altruism

Let us have a closer look at the first stage of Fehr et al.'s experiment. The interaction between B and C is, of course, a classical prisoner's dilemma (PD in the following; cf. Fig. 5.1). Both participants are better off if both decide to transfer. However, each one is better off if she or he does not transfer. If both decided to transfer, each one receives three times the transferred sum. If only one transfers, the party that does not transfer cashes in three times the transferred sum in addition to his own money, which he keeps, while the transferring party ends up with nothing. If neither transfers, both keep their own money.

		A	
		cooperate	*defect*
B	*cooperate*	R/R	S/T
	defect	T/S	P/P

Fig. 5.1 The Prisoner's Dilemma game (if T > R > P > S)

In the classical notation of the PD, the "transfer"-option is labeled "cooperate", the "non-transfer" option "defect". The payoffs are labeled R, S, T and P for "reward", "sucker's payoff", "temptation" and "punishment". Without doubt, the PD is the best analyzed problem in all of game theory. The amount of ink spilled on the issue is truly staggering (cf. Poundstone 1993 for a first overview). And it seems hardly an exaggeration to say that the PD has been the *paradigm* of social theory over the last half of a century, determining the angle from which the phenomenon of the social has been approached. And yet, for all of the attention it has attracted over the last half century, a certain way of interpreting the PD has been so predominant that it is rarely noticed in the current debate that it is an *interpretation* of the PD rather than the PD itself. In this interpretation, the PD illustrates something like the tragedy of economic rationality, which leads to Pareto-inefficient results, or the impossibility of mutually beneficial cooperation among rational egoists. As I shall argue, however, the labels "egoist" and "cooperation", as applied to the PD, depend on a certain interpretation of the social identity of the participating agents, which has consequences for what counts as a solution to the problem.

The conventional interpretation is deeply rooted in the history of the PD. The original design of the game is usually credited to Merrill Flood and Melvin Dresher, whose original idea was given its name and narrative clothing (to be inspected in more detail below) by Albert W. Tucker (Tucker [1950] 1980). Around the same time (early in the year 1950), and without any apparent connection to the

former, Howard Raiffa stumbled upon the PD in his own game theoretical and game experimental research. Raiffa recounts having had no qualms whatsoever calling mutual defection the "solution" to the PD "from a descriptive and prescriptive perspective". According to Raiffa, the whole point of the PD is simply that "two stupid players do better than two smart players".[1] Raiffa's rather hard-nosed attitude to the problem of the PD seems to be typical for what one might label the 'orthodox' view. It is important to notice, however, that in this orthodox view the PD is not a dilemma. It is part of the definition of a practical dilemma that the agent is forced to choose between equally repellent alternatives.[2] For rational, un-sympathetic egoists such as Raiffa's 'smart players', however, there is no pondering over what to choose in the PD. There is one *strictly dominant* strategy, i.e. a strategy that leads to a better result no matter how the other chooses, and whoever is 'smart' enough to let her or his choice be determined by the expected payoff will have to choose accordingly. Within the orthodox framework the PD is no dilemma, for a dilemma requires being torn between two options. For agents whose rationality is determined by strategy dominance, there is no dilemma. Rather, there is is a practical paradox (as it is indeed often referred to in the relevant literature): by each one choosing what is the optimal strategy, independently of the decision of the other and in terms of the participant's utilities, both participants end up worse off. In this situation, rational choice paradoxically turns out to be a rather ineffective means of maximizing one's utility.

Yet for both normative and descriptive reasons, the hard-nosed 'orthodox' attitude to the problem of the PD is met with increasing criticism. As far as the normative dimension is concerned, one might see Raiffa's alleged 'smart players' as mere 'rational fools' (Sen 1977); at least it seems overly harsh (and indeed incompatible with our pre-theoretic understanding of the term "rationality") simply to discard mutual cooperation as 'stupid' and irrational, as Raiffa suggests. And, descriptively, experimental economists have revealed surprisingly high levels of cooperation even in one-shot PDs among anonymous participants (cf. Kagel and Roth 1995: 26ff.). Thus, for less hard-nosed economists and experimental game theorists the decisive question is: how can these cases be accounted for without dismissing the behavior as irrational?

A broader conception of the structure of human motivation seems to offer a solution. In this view, all that is needed to see how rational subjects can find their way out of a PD is to break with the conception that human behavior be narrowly self-regarding. Egoism is seen as the cause of the prisoners' problem, and altruism

[1] Raiffa (1992: 172); Raiffa started having qualms only when he considered finite repetitions of the PD, feeling rather 'dismayed' at the prospect of constant rational non-cooperation (with correspondingly increased costs). To his relief, however, the participants in his informal experiments turned out to be more cooperative-minded (and less rational?) than he had expected.

[2] Homer provides the classic example of a practical dilemma when in the Iliad (IX/410ff.) he has Achilles pondering over his 'twofold fates', i.e. the decision whether to stay in Troy and fight or return home. Fighting, on the one hand, will earn him 'imperishable renown' which means so much in his life – but not to much avail, for his life will then be rather short. Returning home, on the other hand, will considerably prolong his lifespan, but only at the price of his renown which he values so much.

is believed to be the solution. Psychologically speaking (I shall turn to the behavioral viewpoint shortly), people, according to this view, simply have to care about other people's payoffs enough so as not to be tempted to try to get the better of their partners by unilateral defection. However, there are serious doubts in the existing literature as to the range of cases in which altruism is effective in transforming PDs into games where cooperation is the rational choice. It appears that both in the case of sympathetic and self-sacrificing altruism (the first consisting in incorporating the other's utilities into one's own at a certain positive rate, the second consisting in replacing one's own utilities with the other's), cooperation will result only in some special cases. In the other cases, any of the following might happen: either altruism is simply not strong enough (where the rate at which the other's utilities are incorporated in one's own is too low), or altruism leads to indifference between the cooperative and the uncooperative strategies, or the transformation of the game even leads altruistic agents to favor noncompliance, which is likely to undermine cooperative stability (cf. Verbeek 2002: 86–102). In the latter case, a series of new dilemmas might follow from altruistic motivation, where this is common knowledge (cf. Tuomela 2000: Chap. 10). To use Bruno Verbeek's words, altruism does not seem to be the one omnipotent 'cooperative virtue' which it is often claimed to be.

I do not want to go further into the details of this important discussion here, but will instead take another line to cast doubt on the use of a theory of altruistic behavior as a 'solution' to the Prisoner's Paradox. 'Altruism' might be seen as a plausible candidate for a solution of the PD only where 'egoism' is seen as the cause of the problem. This view, however, relies on a certain interpretation of the PD, namely that the parties affected by the outcome of PD-like situations (relative to whom the choices can be labeled as 'altruistic' or 'egoistic') are identical with those who take an active part in it. In most real-life PDs, however, this will not be the case. What is more, this condition is not even met by what might be called the original PD.

Remember the story Tucker invented to illustrate the problem (the original version, together with what might be the first notation of the PD beyond Dresher's blackboard, is reprinted in Tucker 1980: 101): the two players are introduced as a team of "two men, charged with a joint violation of law", and "held separately by the police". They are presented with the well-known deal. Separately, they face the decision whether to confess (and thereby to implicate the other). However, this is not all: interestingly, there is a third party present in Tucker's story, whose role is all but forgotten in later accounts of the PD. It is the State who – for obvious reasons – has rather high stakes in the matter. Even though the State, as Tucker observes, "exercises no choice" in the PD, it "receives payoffs" (Tucker 1980: 101). From the point of view of the State (i.e. the general public which is represented by the state), the Nash equilibrium – mutual confession – is the optimal outcome (the public has a vital interest in a high crime detection rate). In Tucker's payoff matrix, unilateral confession of either prisoner is the second best outcome,[3] whereas mutual

[3] It seems that this is dictated by Tucker's desire to have the PD transformed into a zero-sum game; in this respect, the role of the State in Tucker's PD might be seen as that of a deus ex machina.

non-confession is the worst outcome from the perspective of the public (the crime remains unpunished, crime detection rate is lowered).

The presence of this third party substantially alters the situation at hand. The prisoners' choice is not just whether to cooperate with each other or defect; it is whether to collaborate with each other *or to cooperate with the third*, i.e. the State, or the public.[4] In the usual game theoretical notation of the situation, the third party is simply left out. Only the two prisoners, their available strategies, and their respective payoffs matter. In this limited view, it might indeed appear as if "confess" was the egoistic choice, "not confess" being the altruistic alternative. In a wider perspective that includes third parties, however, the simple conceptual distinction between egoism and altruism is of little use. If the two prisoners choosing to confess and to implicate each other are traditionally seen as rational egoists, one could equally well interpret them as rational altruists. The reason is obvious: while the two prisoners, by mutually confessing and implicating each other, fail to further their respective self-interests, they are quite effective in furthering the interests of the wider public by contributing to a high crime detection rate.[5]

It might be a little hard to believe, though, that their concern for the public interest is what moves the prisoners when they choose to confess (though some such cases have been reported). Yet we do not need to resort to any such motivational story if we restrict the use of the term altruism to its behavioral meaning, as is the case in Fehr et al.'s analysis of the nature of altruism. Here, as in the entire debate on the interpretation of Fehr's experimental games, "altruism" is defined as a costly act that confers benefits on other individuals, regardless of the psychological background (cf. e.g. Fehr and Fischbacher 2003: 785). Even in this behavioral sense, however, mutual defection in PDs, where third parties are negatively affected by cooperative outcomes, shows all the marks of genuine altruism: while being costly to the agents (cooperation would have left both participants better off!), it confers benefits on other individuals (i.e. the general public).

Thus the conceptual distinction between egoism and altruism is unhelpful when it comes to describing agent's choices in PDs, where third parties are affected. In particular, this concerns cases where mutually defective or cooperative outcomes result under conditions where the following conditions hold:

1. The outcome resulting from mutual individual expected utility maximization is Pareto sub-optimal with regard to s.
2. The agents are members of s.
3. The outcome which is Pareto-optimal for s leaves a third party t worse off than the outcome resulting from mutual individual expected utility maximization.

[4] One might, of course, quarrel over which alternative should be labeled 'cooperation', and indeed this is precisely what this paper is about: a matter of the determination of the social identity of the participants in question.

[5] Apparently, without knowing that this three-party setting was already part of Tucker's original conception of the PD, Elizabeth Anderson emphasized this point (Anderson 2001).

As far as mutual defection is concerned, the above considerations have already indicated that, to some degree, it might be left to one's disposal whether one likes to call the respective defective choices egoistic or altruistic: they can be both. Even more interesting is the case of mutual cooperation, for the relevant behavior seems to be neither of the two. Whereas in this (as in any other) case, cooperation clearly shows all the marks of altruism with regard to the other members of *s* – it is costly to the agent (who foregoes the benefit of unilateral defection), and it confers benefits on the other (sparing her the fate of being made the 'sucker') – it is much more difficult to account for the members of *t* in this story. From their perspective the distinction between egoism and altruism simply collapses. With regard to those unfortunate third parties, cooperation is definitely not altruistic, for it leaves them worse off. But it is clearly not egoistic either, properly speaking, for there was an alternative that would have left both the agent and the members of *t* better off (i.e. defection at the expense of the other members of *s*).

The conceptual confusion that results from a wider view extending beyond the participating parties in a PD calls for clarification. The distinction between egoism and altruism does not work here. Apparently, it is not enough to go beyond the traditional selfish model of human behavior by allowing for altruistic behavior; in order to do justice to those pervasive cases of non-selfish behavior, where an optimal outcome for *s* inflicts losses on *t*, more structure has to be added to the behavioral picture. In the current debate, terms like "group-directedness" are introduced for such purposes (cf. Tuomela 2000). In spite of its somber past, I cannot resist the temptation to use a neologism re-invented by the Spanish philosopher José Ortega y Gasset, who in his book on *Man and People* introduced the term "nostrism" or "nostristic attitude" (Ortega y Gasset 1957: 150). He re-invented this neologism (unknown to Ortega y Gasset, it had been introduced in the 1930s by a Nazi philosopher) because he understood the need to go beyond egoism and altruism in order to capture the sense in which much of our behavior is structured. 'Nostristic behavior' is neither self-directed (or egoistic) nor other-directed (or altruistic) but oriented towards our shared goals and concerns.

§18 The Role of Social Identities in Cooperation

The claim that the analysis of group-related behavior requires more conceptual tools than the conceptual distinction between egoism and altruism is not uncontroversial. Elliot Sober's and David S. Wilson's position on the structure of unselfish behavior, for instance, does not imply any such irreducible concept. Even though, in the second part of their seminal book, they propose a thoroughly 'nostristic' reading of non-selfish motivation (in accordance to which "the 'I' is defined by relating it to a 'we"'; Sober and Wilson 1998: 233), they seem to take group-directed behavior to be a mix of egoism and altruism rather than a third, altogether different type of attitude.

Yet the alternative between reducing group-directed attitudes to a mix of egoism and altruism on the one hand, and introducing nostrism as an independent third type on the other hand is not exhaustive. Perhaps nostrism is neither of the two but rather a general structure of which (a) egoism is a marginal case and (b) altruism is an internal feature (for all cases with the exception of (a)). This can be explained as follows:

(a) Nostrism becomes egoism to the degree that *s* is shrunk so as to contain only the agent himself
(b) Nostrism implies an altruistic attitude towards the other members of *s*

The analysis of the mode of reasoning underlying the nostristic attitude and its implications for our understanding of the structure of cooperation have become the center of an extended debate in the last 2 decades (cf. esp. the work of Tuomela 1995, 2000, as well as, among others, Gilbert 1989; Sugden 1993, 2000; Gold (ed.) 2005). Other lines of thought such as Amartya Sen's concern with the structure of committed action and the role of identity in his critique of rational choice theory fit seamlessly in this general venture (Chapter 6 below; Peter and Schmid [eds.] 2007). Most of the work carried out in this context does not, however, directly pertain to the question of altruism as it arises in experimental game theory, since it is aimed at exploring the reasons, motives and intentions of the agents rather than giving functional explanations. However, no functional explanation of group-related behavior can remain completely indifferent to the question of the possible 'proximate explanations' of such behavior, or ways in which the function in question is actually realized. The question arises as to which motives, preferences, and modes of reasoning can be interpreted as the most likely candidates for having evolved to sustain the respective behavior. For beings whose behavior is not exclusively prompted by instincts and immediate urges but who can think and deliberate, theories such as those of we-mode thinking (Tuomela), team reasoning (Sugden), collective intentionality (Searle), and joint commitment (Gilbert) offer the most plausible candidates.[6]

It is an open question whether, by means of some utility transformation rule, nostristic agents' choices can be fitted into the classical game theoretical framework. I do not intend to pursue this general issue here[7] but limit myself to addressing the

[6] Antti Saaristo argues that, at least where the evolutionary basis is conceived of in group selectionist terms (Sober and Wilson 1998), collective intentionality is the most likely candidate for a proximate explanation (Saaristo 2005: Chap. 2). Fehr et al. claim that group selection is not necessary for the evolution of strong reciprocity (Bowles and Gintis 2003).

[7] In the respective debate there are several attempts at formalizing the group-oriented point of view within the classical game theoretical framework, e.g. by applying transformation rules (for a detailed discussion cf. Tuomela 2000: Chap. 10). It seems to me, however, that Tuomela (2000); Anderson (2001); and Hurley (1989); as well as many others, are right to say that there might be a systematic barrier to any such attempt. It could be this: the game theoretic framework imposes an act consequentialist understanding of choice (where choices are understood as the causes of the outcomes). It appears, however, that if people contribute to shared practices, they conceive of their individual choices not just in terms of cause and effect but in terms of parts and wholes, too. People

special question of how the presence of nostristic agents alters experimental PDs in a non-formal way. Let us define an agent's identity simply as her membership in (or belonging to) the group to whose optimal collective choice her individual choice contributes. Thus, an individual utility maximizer's identity consists simply in being herself, while a cooperating prisoner's identity consists in his being a part of the team of two criminals. How do these identities alter the PD? As mentioned above, the prisoner's dilemma is no dilemma for agents whose identity is limited to their individual selves but a practical paradox (by doing their best to maximize their individual utilities they end up being worse off). One might think this paradox does not arise for team members. As it turns out, however, this is not necessarily the case: agents whose identity extends to the group of their possible co-operators may face no less of a paradox than the individual selves. The problem is the following: as unconditional cooperators, hard-nosed team thinkers will inevitably attract free-riders (who can always count on unconditional cooperators). In many cases this will result in a paradoxical effect. For depending on the circumstances it might well be that unilateral cooperation resulting from the team-thinker's being abused by free-riders is even worse an outcome from a group perspective than mutual defection. Where there is common knowledge of this structure but not of the participant's identity, the PD becomes a paradox even for hard-nosed team-players.[8]

Thus the Prisoner's Dilemma is a practical paradox not only for unconditional defectors (i.e. individual utility maximizers), but for unconditional cooperators (i.e. hard-nosed team players), too. For neither of them, however, is there anything like a dilemma in this situation. Neither *homines oeconomici* nor their cooperative-minded counterparts, *homines sociologici*, will seriously be able to take both possible choices into consideration. For an individual utility maximizer, 'defect' is just as obvious a choice as 'cooperate' is for such 'over-socialized' agents as fully class-conscious workers or citizens of the ideal 'kingdom of ends'. Both images of human behavior are similarly askew. We need to go beyond *both* types of a priori fixation

who choose to cooperate in one-shot prisoner's dilemma situations choose to do their part in what is better for both. Their choices are intended as contributions constituting an optimal collective choice (which can then in turn be evaluated in act consequentialist terms). Thus it seems that a constitutive understanding of choice is required which defies the framework of classical decision theory.

[8] Thus one might imagine that in some case similar to the original story of the Prisoner's Dilemma it comes as quite a shock to Row who has chosen to confess and thereby to implicate her companion when she learns that for his part Column had decided not to confess and thus not to implicate Row (meaning that Row now gets off scot-free while Column faces an extra long time in prison). It might well be that Row is not pleased with this result at all; she had decided to defect not because she was particularly keen on getting the better of Column. Rather, Row had reasoned as follows: while she never suspected Column of trying to get out of the situation by unilateral defection (because she knew Column to be too much of a team-player for that), Row felt certain that Column could not possibly have enough confidence in her so as to choose to cooperate (which, as she is later to find out to her great distress, was a wrong belief). Because she was not particularly keen on ending up the 'sucker' herself, and expected something similar to be true of Column, she decided to defect, and expected Column to defect, too. In this situation, Column's unflinching team-mindedness paradoxically turns out to lead to an outcome that is worse for both.

of human identity in order to understand what is so gripping about the Prisoner's Dilemma. It is a practical dilemma only for those agents whose identity *is not fixed either to themselves or to a given group*. It is a real practical problem only for those agents who can (and have to!) *determine their identity* by making their choices.[9]

Agents whose identity is not fixed a priori can 'see' the situation at hand both as individual utility maximizers and as team-players, and where necessary they will make the task of having to determine their identity by making their choices a part of their reasoning. Even in a one-shot PD situation, Row (who is in a PD with Column) might start out her chain of reasoning by conceiving the situation at hand from the perspective of her individual viewpoint. Given the paradoxical effect of individual utility maximization, it soon becomes evident to her that it would be much better for both to frame the situation as a team. As soon as she re-conceives the situation as 'one of us', however, it immediately strikes her that, given the strong pressure for unilateral defection, it is very likely that her decision to cooperate will lead to an outcome that not only leaves her with the sucker's payoff but seems even worse than mutual defection from the team perspective. Thus it becomes attractive even from a team perspective to adopt the 'each one for himself' approach again and mind one's own utility, looking at the situation not as a group member but as an individual. This leads her back to the beginning of her chain of reasoning. Oscillating between her identities as an individual and as a member of the team, Row has to choose between two equally paradoxical alternatives. Having to determine their identity (i.e. having to choose the 'unit of optimization', as it were) is the prisoner's real dilemma.

§19 'Nostrism'

What is the importance of these considerations for the interpretation of the results of Fehr et al.'s third party punishment-experiment? At first, the connection might not seem obvious. Compared to Row and Column in the above case, the participants in the first round of Fehr et al.'s experiment (B and C) found themselves in a much easier position. Apparently, they did not have to go through any such identity-shaking considerations, for they knew the Third Party (above: the test person A) to be watching them and to exert his influence in the experiment, and it seems that they quite correctly suspected that any unilateral defection from their part would trigger the third party's wrath (in similar experiments, surveys revealed that the participants expected the third party to 'punish' unilateral defection even more severely than he actually did). In a faint analogy to Jean-Paul Sartre's social ontology, where the third party's view 'glues together' the interacting 'I's' to a 'We', the third party here makes cooperation a more rational choice. This is certainly the case for those conditional cooperators whose identity is not fixed to themselves and who are therefore

[9] It is most important not to conceive of the determination of the agent's identity in terms of a rational choice, for this immediately sets off an infinite regress (cf. the 'priority of identity to rational principle' in Anderson 2001).

willing to do their part in what is best for both if the other one follows suit. More than that, the third party even seems to make any understanding of the two as members of a team superfluous by bringing their individual self-interest in harmony with what is best for both. When a sufficient number of third parties can be counted on being around, even simple textbook *homines oeconomici* with their identities limited a priori to themselves will cooperate. Thus the result of the experiment seems to be that the homo oeconomicus model of human behavior and its atomistic view of the agent's identity can be left intact, only that the 'police' has to be added to the picture. Where third party punishers are around, cooperative behavior is simply individual utility maximization under conditions where social norms are enforced (by sanctioning deviant behavior). No further considerations concerning the agent's identities (we-mode thinking, team reasoning, collective commitments and alike) are necessary.

I think, however, that any such reading of Fehr et al.'s results would be profoundly mistaken. A conception of social identity in terms of team membership and collective agency is needed if we are to make sense of the third party's behavior, and of the cooperative norms she enforces. The core idea is that, in more complex settings, cooperation and social identity are mutually explicative. If one labels a certain behavior as 'cooperative' (or 'altruistic'), this is meaningful only with regard to a certain limited set of people, whereas one can always find some other set of people with regard to whom the same behavior would have to be called defective (or 'egoistic'), and *vice versa*. Thus in so far as the third party is interpreted to sanction defection or to enforce cooperation, the problem concerning the determination of the participants' identity in a PD by no means becomes an obsolete issue. Rather, it is transformed into the question of how the identity of the participants is determined by the third. Had she chosen the relevant social identity to be that of the tax payers, the third party would probably have sanctioned the appropriation of the experimental fund for private benefit which resulted from mutual transfer in the first round of the experiment.

In order to make sense of the third party's actual behavior as recorded by Fehr's experiment, one has to assume that, in his perception of the situation at hand, the relevant social identity was the team of the two participants of the first round.[10]

It is a well-known fact that 'shared identities' are of great influence on cooperation in social dilemma situations. In social identity theory as well as in other research programs, experimental studies have repeatedly shown that cooperation rates between members of the same group are much higher than between members of different (or even competing) groups, where the participants know about their partner's social identity (cf. e.g. Kollock 1998). But how do social identities arise? How are social identities determined? How does the perception of the situation as

[10] Remember that the concept of identity, as introduced above, is not a 'thick' concept. In this sense, identity does not necessarily involve such elements as a shared history and a common perception of the situation. In this sense, the total anonymity of the experimental situation does not render impossible the emergence of shared identities.

a member of one or another group become salient to the participants in the experiment, especially where anonymity is part of the experimental setting?

In the case of the experimental game at hand it seems plausible that the *instruction given to the participants* by the experimenters might have played an important role.

That the 'principle of description invariance' (according to which it should not matter to the outcome how an experiment is described) does not hold is no news (Tversky and Kahnemann 1986). Camerer and Fehr, however, suggest that in their experimental games such effects are minimized by avoiding 'concrete' descriptions. To quote their own words:

> The games are usually described in plain, abstract language, using letters or numbers to represent strategies rather than concrete descriptions like 'helping to clean up the park' or 'trusting somebody in a faraway place'. As with other design features, abstract language is used not because it is lifelike, but as a benchmark against which the effects of more concrete descriptions can be measured. (Camerer and Fehr 2004: 58)

How are these principles put into practice? In order to see how this is done, it is necessary to have a closer look at the instructions given to the participants of the third party punishment-experiment.[11] This immediately reveals that at least as far as the question of social identity is concerned, the description of the experiment was much more concrete than the above quotation seems to suggest. Firstly, the two parties in round one are explicitly introduced as a 'group' in the instructions, and they are repeatedly referred to as 'team members' throughout the instructions. In line with this labeling, the possible results of the first round of the game are explained in a list with the respective payoffs labeled 'your income' and 'income of the other member of your group', respectively. These instructions were known to the third party (who had been actively involved in the first round of an experiment as a member of another group). Thus it seems quite understandable indeed that she came to determine the team of the two (rather than, say, the taxpayers, or a team consisting of herself and the experimenters) as the relevant social identity underlying the 'cooperative norm' which she decided to enforce in the second round.

With this in mind, let us come back to the question of how the "sketch" provided by the experiment should be interpreted. How should the third party's behavior be labeled? In behavioral terms it is not egoistic or self-directed. Is it therefore other-directed, as Fehr et al.'s label 'altruistic' suggests? Should the result of the experiment be taken to show that human beings do not like having defectors around (or rather: like to inflict losses on them wherever they find them), regardless of how the social identity relative to which the behavior in question appears as 'defection' relates to their own social identity? I think that this is rather implausible, and given what a look at the instruction revealed, it is clearly not the case in the experiment at hand. The set of people whose total payoff the third party's behavior optimized was the group of all participants of the experiment, including herself. Again, this

[11] I wish to thank Ernst Fehr and Urs Fischbacher for generously giving me access to this unpublished material.

particular social identity is heavily supported by the instructions given by the experimenters; the label “participant” (which is repeatedly used in the instructions) alone makes the total set of interacting individuals salient in terms of social identity. In addition to that the participants are told that they are taking part in an experiment. The term experiment – and even more clearly the term experimental game – again suggests and supports an understanding of the situation at hand according to which no ‘outside relations’ matter. Just like a game, an experiment presupposes some sort of isolation from outside influence. Thus individuals who are told that they are participants in the experiment will almost inevitably see themselves as ‘one of the participants’ rather than, say, family members, or taxpayers. Thus, the third party’s behavior should be seen as a sign of a ‘nostristic’ attitude as one of the participants of the experiment, i.e. a personal investment in the maintenance of the ‘normative infrastructure’ that is ‘best for us’ in terms of the total group of the participants in the experiment.

This nostristic reading of Fehr et al.’s experiment does not conflict with Fehr et al.’s own proximate explanation, according to which A’s behavior is driven by his emotional response to the observed behavior. It is a well-known fact that our emotional responses are heavily influenced by our perception of the social identities of the participating parties. We tend to respond in different ways to cases of unilateral defection depending on whether or not the defecting party or his ‘victim’ is ‘one of us’. Nostrism is not just a matter of cognition; it is a matter of affection, too. More than that, the nostristic reading of Fehr et al.’s results seem to be in tune with most of the labels which Fehr et al. attach to their results. If we conceive of the social identity of the group of participants as relevant in the situation, most of Fehr et al.’s abovementioned labels make sense. The third party’s behavior does indeed appear to be ‘pro-social’ and ‘norm enforcing’, as Fehr et al. claim – even though these terms should be handled with care because they only make sense on the basis of a prior correct identification of the relevant social identity. Something similar is true for the label “reciprocity”. In terms of personal interaction, the view put forward above in §15 might be correct: because of the structure of the experiment there was no reciprocity whatsoever involved. However, it is not absurd to call the third party’s behavior reciprocal in the sense that she is doing her part in a communal cooperative practice. The reciprocal element in her behavior consists in the fact that she sanctions defective behavior in the team over which she watches just as she can count on being watched by another ‘third party’ in her own first round of the experiment. The fact that the person she can punish is not identical with the person by whom she can be punished herself does not mean that the attitude is not one of ‘reciprocation’ in terms of doing her part within the grid of interchangeable roles. One might still quarrel over whether or not ‘punishment’ is a lucky terminological choice for the behavior at hand. In any case: in order to make good sense of the experiment, it is essential to understand the social identities of the participating agents.

The experimental games show how little it takes for people to come to conceive of themselves (and act) as members of a team. In view of the result of experimental economics, the Aristotelian dictum of the zoon politikon takes on a new meaning: even with complete strangers whose names and personal identities they do not know,

and even in pure one-shot interactions, people tend to see themselves (and conceive of others) as members of teams. Contrary to what one of the fathers of the PD (Flood 1958: 12) has conjectured, cooperation seems to be much less a matter of some social relationship between the players (there is none), than a matter of their social identity.

Thus, it appears that the lesson to be learned from game experiments is not just how far the orthodox economic assumption of narrow self-interest is from reality. Such experiments shake yet another pillar of economic theory, the last pillar that has largely survived the recent boom in reconsideration and reconceptualization of the economic model of human behavior. What is at stake here is the methodologically individualist view that social phenomena should ultimately be explained exclusively in terms of individual action, without there being any collectivity concepts involved at the basic level of explanation. In cases such as the one at hand, however, the agent's identity has to be determined before it is even possible to make sense of the observed behavior. Thus there is no 'understanding' of the behavior of individuals without first determining the relevant groups of which these individuals are members. Thus groups are not 'secondary' to (or 'supervenient' on) individual action but an essential feature of the most basic level of social reality.

Chapter 6
Rationalizing Coordination

The aim of this chapter is twofold. First, the focus is shifted from *cooperation* (as discussed in the previous chapter) to *coordination.* The theory of coordination provides another excellent example for the importance of the analysis of collective intentionality to the social sciences in general and economic theory in particular. In contrast to the last chapter, the issue at stake here is purely theoretic (§20 below). Coordination problems differ from *cooperation* problems in that, intuitively speaking, rational participants aim at "matching" their individual choices among the available alternatives, so that there is no incentive for unilateral defection. Whereas the structure of cooperation (as exemplified by the prisoner's dilemma) has been at the very center of much thorough philosophical analysis – as well as of experimental work in economics – over the last half-century, the structure of *coordination* has not received nearly as much attention. One reason for this is that in most real-life cases coordination problems are easily solved by means of *conventions*. But coordination is a problem for orthodox rational choice theory. In line with some recent literature on the topic, I will argue that, because of its individualistic limitations, the standard economic model of human behavior fails to explain how conventions make coordination among rational agents possible.

In the second part of the chapter (§§21–22), it is argued that even though the existing accounts of collective intentionality point the way towards an adequate account of coordination, a stronger conception of collective intentionality than the ones to be found in the existing literature is needed. In a discussion of Robert Sugden's theory of team thinking, some features of an adequate account of collective intentionality are introduced.

§20 A Philosophical Scandal

For all its blatant absurdity, the following fictional story might serve as an introduction to the problem. On a sunny afternoon, the police are called to the site of an accident. A look at the scene immediately makes clear what has happened. On a street with separate lanes, a car has crossed the middle line, thus coming into the oncoming traffic's way, eventually resulting in a head-on collision. Luckily, no one

H.B. Schmid, *Plural Action: Essays in Philosophy and Social Science*, Contributions to Phenomenology 58, © Springer Science+Business Media B.V. 2009

is seriously hurt. The police question the culpable driver. Why did he cross the line? Did he lose control of his car, did he experience any technical difficulties prior to the accident, or was he inattentive or somehow distracted? The driver answers in the negative: he had become aware of the oncoming car early enough, he was in full control of his car at all times, and he knows the traffic rules, which, as he admits, were in no way violated by the other driver. The officer asks him if he wanted to kill himself. Again, he answers in the negative. He has, he claims, no suicidal or otherwise destructive impulses or desires, nor any reason to assume that the other driver might have any such preferences. So why, then, did he end up on the wrong side of the street? He answers with a stern expression on his face: "I just couldn't see why I should keep to my own side of the street rather than swerve to the other's in order to avoid a collision."

For a moment, the police officer in charge loses his straight face in disbelief, and a disparaging remark concerning the driver's state of mind slips from his mouth. Now it is our driver's turn to lose his calm. Angrily and firmly, he asserts that he knows perfectly well that he is the guilty party in terms of the law, and he will accept all charges in terms of legal accountability, but he rejects any accusations of irrationality. He explains that, in hindsight, he knows very well that he would have better chosen to stay on the right side of the street. But that does not mean, he claims, that his decision to swerve to the other side was *irrational* at the time it was made. For when the oncoming car approached, he had to decide between the two strategies "right" and "left." As opposed to right, left is illegal and therefore is connected with the risk of getting fined by the police should the encounter result in a collision. At the same time, however, it would be plainly irrational *not* to commit a minor traffic violation such as choosing "left" if it is the only way to avoid an accident.

"Thus I knew," our driver concludes, "that if the other driver had decided to swerve to the left, it would have been rational for me to swerve to the left, too. And since all of this is common knowledge, and since the other driver is a rational being too, I assumed that he must be having the same thoughts, thinking that I have them too, and so on and so forth. It dawned on me that, however deep my (and his) analysis of the situation would dig, the *rational* thing to do in our respective shoes would always remain *hypothetical*. In such situations, you can say what's rational only *if* you have some expectation concerning the other's decision. But you cannot form such expectations because at the same time, you know that the other's decision in turn depends on his expectation concerning your own choice. Thus, in my reasoning I got stuck in a circle of "ifs," and when we finally reached the point where each of us had to make his final decision over our strategies, I saw there was no way to derive from all that what was *in fact* rational for the other driver and for me to do. You can't say the other driver's choosing right was rational and my choosing left irrational. For, as it turned out, his decision to stick to the rules turned out to be wrong, too, since he could easily have avoided getting stuck in this car accident by swerving to the left!"

Of course there is something wrong with this claim. This whole way of reasoning is profoundly mistaken. From a commonsensical understanding of rationality at least, we should take the police officer's side for once and admit that he is probably

right when he doubts our driver's rationality. To the pre-theoretical eye, it is obvious that it is *plainly irrational* not to keep to the right where the right-side rule applies, if one assumes that the rules as well as the absence of suicidal preferences are common knowledge. But is this strong (and presumably reasonable) pre-theoretical intuition backed by our standard theory of rationality in action, such as is implied in the economic model of behavior? Can our driver be proven wrong within the theoretical framework of individual expected utility maximization?

Our driver is, of course, well aware that there are not only two coordination equilibria in the game at hand (right/right and left/left), but also a *convention*. In David Lewis's (1969) terms, right/right is *salient* or, in Thomas C. Schelling's (1960) terms, a *focal point*, which makes, in a sense, right the obvious choice for each. Our driver does not ignore the existence of the right-side traffic rule, or deny that this rule is common knowledge among the participants. Also, he might easily grant that it is clear from previous experience that the probability of the strategy "right" being chosen by the drivers in a given population is extremely high. What he is getting at is that mere objective behavioral regularities and the existence of precedents do not provide a *reason* for a rational choice. Our driver's point is that he could not form any prior expectation concerning the other's behavior because he knew that as a rational being the other driver would have to base his decision on his expectation concerning our driver's behavior, and not just blindly follow some behavioral pattern. And indeed it's hard to see why it should be rational not to treat the other as a rational being, but as some sort of compulsive salience seeker (we will come back to this later). If, however, the other party is treated as a rational being, it is clear that it is rational for him to conform to the general pattern only if he expects that the same will hold true for our driver himself, which sets off the infinite regress of interdependent expectations.[1] In Raimo Tuomela's words, what our driver is getting at is the "deconditionalization problem" in the theory of coordination (Tuomela 2002a: 388). "Deconditionalization" here means getting rid of the condition on which deriving on which "right" seems to be the rational choice, that is, the *Big 'If'*: "right" is the rational choice *if* one has no reason to expect the other to choose "left." It seems impossible, however, to derive from the hypothetical (or "conditional") rationality of "right" that, given common knowledge of the right-side traffic rule and the absence of suicidal preferences, "right" is *in fact* the rational choice.[2]

Deconditionalization can be a real-life problem in situations where there are no conventions, and no salient solution or focal points (as in the case of two pedestrians colliding on the sidewalk because of a mismatch of the chosen strategies). Where there are commonly known conventions (such as in the case of motorized traffic), however, these problems do not usually occur. It seems that in these latter examples

[1] For a very clear formulation of the problem, compare Parsons and Shils (1951: 105).

[2] "An agent cannot rationally [. . .] form and satisfy his action intention without a circular reference to the other agent's intention. Hence, he cannot, so to speak, finitarily infer or compute the satisfaction value of statements like 'I will do X if you will do X' in the kind of coordination situation on the basis of independently assignable satisfaction values of 'I will do X' and 'you will do X,' because there simply are no such satisfaction values" (Tuomela 2002: 390).

deconditionalization is easily achieved. We do not experience serious difficulties in deciding whether or not to stick to the traffic rules in order to avoid a collision, where the rules and the absence of suicidal preferences are common knowledge. Our pre-theoretical intuition is that, in a very basic sense, this is simply a matter of *rationality* in terms of straight reasoning. One of the philosophical questions behind this is the following: How precisely do salience and the existence of conventions provide us with reasons? Why is it rational to choose the salient strategy in pure coordination games, when apparently the existence of conventions, focal points, or salient solutions do not *per se* solve the deconditionalization problem? Or, put negatively: what precisely is wrong about our driver's claim that, right-side traffic rule or not, any attempt to base a rational decision on salience is immediately drawn back into the infinite regress of interdependent expectations?

This problem is a serious one; considering the importance of orthodox accounts of practical reasoning in contemporary social science, it seems appropriate to call the deconditionalizing problem a *philosophical scandal.* If orthodox rational choice theory cannot endorse our pre-theoretic view that it is rational to stick to conventions where pure coordination situations are concerned, and the situation as well as the rationality are common knowledge, we have to accept the fact that the most important theory of practical reasoning flies in the face of our most basic intuitive understanding of what it is rational to do in social situations. How, then, could we expect any real understanding of social reality to come from social science that is based on such assumptions about practical reasons?

It should be noted in passing that not all social scientists have been ignorant of this problem. Indeed it seems that Max Weber, Talcott Parsons and Niklas Luhmann have struggled with this issue, the latter two under the label "multiple contingency". Both Parsons and Luhmann took the problem to be so serious as to demand a deep-reaching revision of the conceptual tools to be used in social science, turning away from intentionalism and action theory towards systems theory (I shall turn to their solutions in Chapter 12 below). I will argue that it is wrong to draw such conclusions. The problem isn't intentionalism or action theory; rather, it is the individualistic way in which intentionality and action have been conceived so far.

But let us now see how the problem has been dealt with within the "Humean" model of practical reason.

§21 The Principle of Coordination

In his very influential paper on the topic, David Gauthier (1975) tried to answer this question. Going beyond Lewis and Schelling, his brilliant move was to draw attention to the role of the *description* under which the players perceive their available strategies.[3] For drivers who do not know any conventions, the alternatives at hand are simply "right" and "left."

[3] For a more detailed account of Gauthier's approach, compare Sugden (1995).

Fig. 6.1 A pure coordination game

		A	
		left	*right*
B	*left*	0/0	–1/–1
	right	–1/–1	0/0

The situation might look like Fig. 6.1. For rational players ("rational" in terms of individual expected utility maximization), it is clear that, according to the "principle of insufficient reason," they will choose randomly between the two strategies, which makes all four possible outcomes equally probable. If, however, "right" is salient, each player's choice is now a different one: it is between the alternatives "choose salience" (that is, choose right) and "ignore salience" (that is, choose randomly). This choice has a single coordination equilibrium, viz. both choose salience.

After the "relabeling" of the available strategies, the situation looks something like Fig. 6.2. Thus, by means of salience, a game with two equally good coordination equilibria is miraculously transformed into a game with only one best coordination equilibrium, which makes it rational to choose the corresponding strategy, that is, to stay on one's own side of the street.

	choose salience	ignore salience
choose salience	0/0	–0.5/–0.5
ignore salience	–0.5/–0.5	–0.5/–0.5

Fig. 6.2 Pure coordination, re-labeled strategies

All of a sudden, the deconditionalizing problem seems to disappear. However, this transformation is too good to be true, and our driver will have no difficulty proving Gauthier wrong (cf. Provis 1977; Miller 1991; Gilbert 1996; Goyal and Janssen 1996). For Gauthier simply ignored that in the transformed version of the game the choice is not between *two* strategies, but between *three*. The options open to the players are not just *either* to observe salience (i.e. choose right) *or* to ignore salience (i.e. choose randomly); the third option is to *choose the non-salient* (i.e. to choose left).[4] In other words: in a derivative sense, the existence of *salience* makes the non-salient strategy salient (one might speak of *secondary* or *derivative* salience). This third strategy, however, has another equally good coordination equilibrium (both choose the non-salient), which throws us right back into the initial deconditionalizing problem. The "re-labeled" situation is shown in Fig. 6.3.

Another line of argument in Gauthier's paper goes as follows. The drivers could get by each other either by both choosing "right" or both choosing "left". However,

[4] Provis has a convincing explanation of why this obvious weakness of his transformation argument could have slipped Gauthier's notice: the existence of the third option "is obscured because Gauthier introduces his suggested alternatives as being choosing the salient option and ignoring salience. That phraseology diverts attention from the fact that one way of not ignoring salience on an option is by performing the non-salient option qua non-salient option" (Provis 1977: 509).

	observe salience	*ignore salience*	*choose the non-salient*
observe salience	0/0	–0.5/–0.5	–1/–1
ignore salience	–0.5/–0.5	–0.5/–0.5	–0.5/–0.5
choose the non-salient	–1/–1	–0.5/–0.5	0/0

Fig. 6.3 Pure coordination, re-labeled, complete set of strategies

right/right is not only the *salient* solution. It is also *payoff dominant*, that is, better for both, for by choosing left both drivers run the risk of getting fined for violation of the traffic regulations even if they manage to get by each other collision free. (The risk of getting caught in the act is low, but it rises to certainty when a collision results, for then the police will be called in and the culpable driver will be fined.) Thus the two coordination equilibria in the initial game are not equally good after all, which means that right is *weakly dominant* for both.

Thus a more adequate description might reveal an asymmetry between the two coordination equilibria in the original situation, as shown in Fig. 6.4.

	right	left
right	0/0	–1/–2
left	–2/–1	–0.1/–0.1

Fig. 6.4 Coordination, unequal equilibria

Compared to right/right, left/left is Pareto inferior. As Gauthier states in his famous "Principle of Coordination," this makes it rational to choose the superior strategy. Or, in Gauthier's own words:

> In a situation with one and only one outcome which is [. . .] a best equilibrium, if each person takes every person to be rational and to share a common conception of the situation, it is rational for each person to perform that action which has the best equilibrium as one of its possible outcomes. (Gauthier 1975: 201)

However, it is not as obvious as it might first appear that the principle of coordination is a *rational* principle, and indeed it seems that what our driver is getting at is that it is not. In spite of the fact that left/left is Pareto inferior, it is still an equilibrium, that is, it is rational to choose left if one expects the other to choose left also. Thus it appears that, just as in the case with two equally good equilibria, both strategies are *hypothetically* rational. Or, in other words, it is rational for the single participants to adopt Gauthier's Principle of Coordination only if they expect it to be adopted by the other participant, too. Thus, once more, the participants see themselves confronted with the deconditionalizing problem. From the participants' perspective, the question "Why does rationality require *me* to follow this principle?" remains open (Hollis and Sugden 1993: 11).

This, however, did not slip Gauthier's notice. Against possible objections of this type, he argues that because of the coincidence of the individual payoffs in both equilibria the participating individuals can identify their respective individual choice

over the two possible strategies (right and left) with a *collective choice* over the two equilibria. Since right/right is better for both than left/left, it seems clear that right is the rational individual choice. In a much similar vein, Maarten C. Janssen has argued that right is the rational choice since whoever so chooses picks the "better plan" (Janssen 2001a, b).

Will that finally convince our driver? There is reason for doubt. That left/left is Pareto-inferior as compared to right/right did by no means escape his notice. Thus he will grant that right/right is the better plan. What he points out, however, is the fact that when the oncoming car approached he had to make an *individual* decision, not a common decision. It had nothing to do with a decision over plans; it was a decision between strategies. In his view, this difference is a decisive one, for it is all that his reasoning was about: the strategic interdependence of decisions in the given situation. It is simply impossible to derive from the fact that right/right is the "best plan," in terms of the optimal outcome which would be rationally chosen in a *common* decision, that right is the rational *individual* decision for either driver in the given situation. Whether or not there is a single best equilibrium, as long as there are multiple coordination equilibria, the basic deconditionalizing problem remains the same.

For all that remains for a more detailed analysis of the problem at hand, it might have become apparent how stunningly little there is to say against our driver within the framework of the standard theory of rationality in action. If this should turn out to be true, it seems that what we have here is an open conflict between theory and common sense. If the foregoing is right, this tension is not just an *apparent* tension; there is something wrong either with our pre-theoretical intuition or with the theory.

So what is at fault here – common sense or the standard theory? This is by no means a rhetorical question. Not all philosophers who believe that the standard theory of rationality in action cannot be reconciled with the commonsensical intuition concerning the rationality of coordination hold that the theory is at fault. Quite the contrary: some philosophers state that we should let go of our pre-theoretical intuition rather than revising the standard theory. Famously, Thomas C. Schelling seems to be vaguely pointing in this general direction when he states that success in pure coordination games "depend[s] on imagination more than on logic" (Schelling 1960: 57).

It appears, however, that once we start saying that focal points or salience do not provide a reason for a rational decision in pure coordination games, we end up having to grant our driver that he was right after all in saying that there was ultimately no *rational* decision in his particular situation. Such a concession not only ruffles the feathers of our common sense; it also gives rise to the question: how do we reliably manage to coordinate along the guidelines of the traffic rules if it is not outright *rational* to stick to those rules? Is this system of conventions on which we rely so much in our everyday dealings really built on mere *imagination*?

Those philosophers who take the *irrationalist* position in the current debate on the structure of coordination usually quote some non-rational impulse (for example Thalos 1999), some psychological propensities or "blind" behavior (cf. Gilbert 1996). In this view, our driver was by no means *irrational* when he failed to choose right; he just happened not to act on the usual kind of impulse.

Is this irrationalist position right? I do not think that we have to go much deeper into the details to encounter some of the difficulties into which this line of argument seems to run. It is obvious that non-rational impulses (including habits, psychological propensities, or some such) are neither a *sufficient* nor a *necessary* condition for human coordinative behavior. Any student driver who knows the rules, and knows that the rules are common knowledge, will stick to them, even though she has not yet acquired any corresponding habits or extra-rational propensities and impulses. It is not the corresponding impulses that make it rational to choose the salient strategy; rather, it is because it is rational to choose the salient strategy that we acquire the corresponding impulses and habits. In other words, the irrationalist position puts the cart before the horse.

This becomes even clearer when we consider those (admittedly rare) cases where it is not by acting on impulses but by *suppressing* them that coordination is achieved. As any continental European or American who has ever driven a car in Great Britain or Australia will confirm, it is possible to coordinate along the lines of the left-hand traffic rule in spite of persisting impulses to the contrary. In the case of the people coming from continental Europe and driving off the car ferry in Dover, a great deal of impulse suppression is required. It seems plausible that this is only possible because under normal circumstances (that is, where common knowledge of those rules and the absence of suicidal preference can be assumed) people such as student drivers and foreigners find it perfectly *rational* to stick to the traffic regulations in order to avoid collisions.

Against this, an irrationalist philosopher of coordination might argue the following: the fact that, under special and unusual circumstances, such as learning to drive (or driving according to unfamiliar rules), coordination is achieved rationally without appropriate impulses (or even by suppressing them) does not disprove the fundamental role of impulses in coordination, because it is the *other drivers'* impulses to which it is rational to adapt one's own behavior. Thus it still seems to hold that any sound reasoning about what equilibrium to aim at in coordination games ultimately bottoms out in mere impulses. However, this defense of the irrationalist position does not stand up to closer scrutiny. It is not necessary either to act on impulses or to count on the other's appropriate impulses to achieve coordination. Coordination can even be achieved where both parties act against their impulses, and where this is common knowledge.

The following real-life example may serve as an illustration. The island of Jersey, the largest of the Channel Islands, is popular with tourists, many of whom come over from the continent, since Jersey is only some 14 miles off the French coast. There is public transportation on the island, but in order to avoid the crowd and to reach the most beautiful places many tourists, who come mostly from the continent, either rent a car on the island, or even have their own car brought over by ferry. There are many narrow roads with no separated lanes on Jersey; in order to avoid the branches sticking out of the hedges on both sides, most cars drive in the middle of the road, moving aside (often without reducing their considerable speed) only to let oncoming cars pass by. Because the states of Jersey are part of the United Kingdom, this is done by both swerving to the left. The many drivers who come

from the continent have to suppress their impulses in order to adapt to the left-side traffic rule. What flies directly in the face of the irrationalist position, however, is this: even drivers who are clearly recognizable to each other as coming from the continent by their number plates and by the location of the driver’s seat in their cars coordinate without any difficulty by keeping to their left.

From an irrationalist viewpoint, this must appear like a small miracle. For, clearly, these drivers neither act on impulse nor rely on the impulses of the other drivers (since it is common knowledge between them that both drivers are from the continent). But if it is true that in this particular situation there are indeed no appropriate impulses, habits, psychological propensities and such, how then is coordination possible? The obvious answer is very simple: contrary to what the irrationalists say, common sense is correct. Given common knowledge of the rules and of the absence of suicidal preferences, sticking to the basic traffic rules is simply the *rational* thing to do.

It is true that, as experienced drivers in our everyday world, we do not *think* about whether or not we should stick to the rules; we just do it “blindly.” This, however, does not mean that rationality is not involved here, or that it comes second to our habits and impulses. For if we let ourselves “blindly” be guided by the rules, we do this precisely because we think that this is the rational thing to do. And how could this belief be so pervasive if it is wrong?

§22 “Team Thinking”

If the foregoing is right – that is, if we can neither accommodate our pre-theoretic intuitions in our standard theory of rationality in action (§3), nor let go of our pre-theoretic convictions concerning the rationality of coordination (§4) – it seems that the correct position will be the only one that is left: to do something about the theory in order to make it fit our deeply engrained pre-theoretic conviction. If so little can be said against our driver from within the conceptual framework of our standard theoretic understanding of rationality in action, we will have to revise this framework. Obviously, there is something more to rationality in coordination than mere individual expected utility maximization in the sense discussed above.

There are several theories that point the way to go. In their *General Theory of Equilibrium Selection in Games*, John Harsanyi and Robert Selten claim that the principle of coordination – their term is “payoff dominance” – cannot be derived from individual rationality, but implies an altogether different concept, a *collective* concept of rationality:

> Our theory uses two independent, and ostensibly very different, criteria of rationality. One of them, risk dominance, is based on *individual* rationality: it is an extension of Bayesian rationality from one-person decisions to n-person games involving strategic interaction among n players, each of them guided by Bayesian rationality. [. . .] In contrast, payoff dominance is based on *collective* rationality: it is based on the assumption that in the absence of special reasons to the contrary, rational players will choose an equilibrium point yielding all of them higher payoffs, rather than one yielding them lower payoffs. That is to say, it is based

> on the assumption that rational individuals will cooperate in pursuing their common interest if the conditions permit them to do so. (Harsanyi and Selten 1988: 365)[5]

Picking up on Harsanyi and Selten's insight, Robert Sugden has developed his theory of *team thinking* (Sugden 1993, 1996, 2000, 2003; Gold and Sugden 2007). Very roughly, the basic idea of Sugden's theory, as well as Michael Bacharach's 1998, 2006) somewhat related account seems to be the following. The problem with the standard theory is that it conceptually restricts the "units of agency" to single individuals. This leads to an inadequate account of those situations, where we do not reason and act as single isolated individuals, but as *members of teams* instead. "Team membership" is basically meant in the sense of participation in collective action. To understand the structure of team thinking, it is important to see the situation at hand as one of *shared intentionality*.

Here, the recent turn to collective intentionality comes into play. Whereas "classical" philosophy of mind focused exclusively on the analysis of *individual* intentionality, it has become increasingly clear over the last 2 decades that in order to account for the social dimension of human action and cognition, the analysis has to be extended to *shared* intentional states. Based on seminal contributions dating from the 1980s, the analysis of *collective intentionality* has gradually evolved into a distinct field of research. The most important theories of collective intentionality are those by Raimo Tuomela (1995; Tuomela 1988), Margaret Gilbert (1989), John Searle (1995), and Michael Bratman (1999), all of which differ in fundamental ways. Tuomela's account rests on individuals' intentions to do their part, together with a structure of mutual belief. Searle criticizes Tuomela and claims that collective intentions are irreducible to sets of individual intentions. Searle's account rests on intentions of the form "we intend *A*" in the individual minds of the participants. Bratman, for his part, gives an account of collective intentions in terms of interrelations of individual intentions of the form "I intend that we *J*". Gilbert again follows an altogether different line by making *collective commitments* the center of her account of collective intentions. These differences notwithstanding, the importance of collective intentionality analysis for our understanding of both the mind and the social world has been widely recognized in philosophy, as well as in many neighboring disciplines.

Can collective intentionality analysis indeed help us to understand the rationality of coordination? With regard to the example of our drivers, this might seem rather unlikely at first. For obvious reasons, the paradigmatic cases of shared intention are cooperatively loaded cases such as Searle's (1990: 400ff.) example of the joint intention to cook a hollandaise sauce, Gilbert's (1996) example of the joint intention to go for a walk, or Tuomela's (1995: 137–138) example of a group of people joining forces to push a broken-down car. By comparison to such intensely communal

[5] Raimo Tuomela seems to adopt a similar view on the deconditionalization problem. He says that deconditionalization is not a "fully 'r-rational"' procedure (Tuomela 2002: 395–396), where "r-rational" is something like individual instrumental rationality. However, he distinguishes r-rationality from a wider everyday sense of rationality, which seems to include the possibility of rational deconditionalization.

endeavors, it might seem that there is nothing genuinely collective or social about our driver’s intention to avoid bumping into another driver whom he does not even know by sticking to the right-side traffic rule.

The theory of team thinking, however, points out the hidden element of sharedness that is implicit in these cases. If “avoiding a collision” is seen as something the two drivers desire individually, the deconditionalizing problem appears to be insoluble. If we conceive of the participating individuals as acting on their individual desire to get by the other collision free, we immediately find ourselves caught up in the regress of interdependent expectations that our driver pointed out. This is not the case, however, for drivers who are seen as basically sharing the aim to get by each other collision free. Given the fact that what we together intend to do is to get by each other, what is rational for me is to perform my share of what maximizes our shared desire. Thus “right” is not rational because it immediately yields a better outcome (which it does only conditionally – if the other driver were to choose left, my choosing left would be in the best interest not only of myself, but of the team also). It is rational for me to choose right because it is my part in what we should be doing.[6] The fact that individuals can be team members has consequences for what it is rational for them to do. For “one of us,” the decision to move left is plainly irrational. Thus the deconditionalizing problem is solved.

The theory of team thinking requires a theory of rationality, intentionality, and action that is richer than the one that is implicitly adopted in the standard economic model of behavior, because it allows for a sense in which *teams* can be said to have preferences (or even make choices) which are, in a certain sense, *irreducible* to simple individual preferences (or choices) (cf. Sugden 2000). In this view, not all preferences, goals, desires, and other intentional states are individual goals and intentional states. In the case of the driver’s coordination problem, there are not two separate individual goals not to collide; the participants act on a *shared* goal instead (cf. Schmid 2005a). Getting by each other collision free is not anything the single individual drivers want. It is something they want *together*. This desire is *irreducible* in the sense that it is not the case that the drivers share their desire (have a preference for right/right) because they have the appropriate individual desires (that is, an individual preference for right); rather, their individual “contributive” intentions or preferences are *derived* from the shared intention or preference.

Thus in this situation the drivers do not appear as distinct units of agency, but as members of a whole that in a sense appears to be capable of thinking and acting. To capture this trait of team thinking, Sugden invokes a rather strong concept of collectivity. It seems that there are not only individuals at the basic level of explanation

[6] In Susan Hurley’s (1989: 145ff.) view, “right” is rational not because of its causal consequences, but because of its constitutive consequences, that is, because our individual choosing right constitutes the action that is best for both of us. I am somewhat uncomfortable with Hurley’s way of rationalizing coordination, because it seems obvious that it renders unanswerable the question “should I do what constitutes my part in what constitutes the best collective choice, or should I rather do my part in what constitutes the second-best collective choice.” Obviously, it is rational to choose the latter alternative if one expects the other to choose that alternative too. In other words, Hurley’s constitutive rationality does not solve the deconditionalizing problem.

of social phenomena, but also teams, to which, following Sugden, the participating individuals are members "in something like the old sense in which arms and legs are members of the body" (Sugden 1993: 86).

The idea of some irreducible sense of collectivity goes much against the individualistic grain of current social theory and social science. It might even appear that in the theory of team thinking some somber group mind raises its head. Indeed it seems that Sugden himself loses some of his anti-individualistic courage when getting sight of these possible ontological consequences of his theory. What could possibly save us from ending up in a collectivist group mind conception once we start loosening the individualistic restriction of the classical account? Sugden resorts to the following solution. In a rather harsh contrast to his strong concept of membership, he hastens to assert that the existence of the collective depends on the participating individuals thinking of themselves in terms of team members, which conforms to the classical, individualistic, Weber-style view of the social, in which collectivities are "real" only as parts of the contents of the intentional states of individuals.[7]

In a quite similar vein, Michael Bacharach based his theory of *team reasoning* on some "group identification," in which individuals come to take themselves to be members of a team: "in certain circumstances, individuals tend to identify themselves with a group; and a group identification leads them to team-reason" (Bacharach 1998: 132). Besides these two important philosophers, there are other attempts to reconcile the acknowledgement of the role of some kind of "team reasoning" with an individualistic ontology of action. Thus Maarten C.W. Janssen puts forth an even weaker version of the role of collectivity in coordination, replacing "collective rationality" with what he calls "individual team member rationality." Again, the ontological line behind this is stoutly individualistic: "where there is enough information and knowledge about each other, players can consider themselves as a team and think individually what is best for the team and its members" (Janssen 2000: 13).

There are at least two reasons, however, to reject the view that team thinking (by any name mentioned above: collective rationality, team reasoning, or individual team member rationality) depends on some "taking oneself to be a member of the team" from the side of the individual members (or some reflective "group identification" or some such). First, this view seems rather absurd if we consider cases such as our driver. We obviously do not have to "take ourselves to be members of a team" to find it rational to stick to the traffic regulations in order to avoid accidents. If team thinking is at work in these cases (and I believe that it is), the element of collectivity involved here is obviously not a matter of some reflective attitude or belief about oneself, for it seems that, phenomenologically speaking, there are no such attitudes whatsoever involved here. Team thinking does not require that the participating individuals take themselves to be members of the team.[8] Reflective

[7] "A team exists to the extent that its members take themselves to be members of it" (Sugden 2000: 192).

[8] This is obscured in Sugden's account because of his preoccupation with more cooperatively loaded cases such as his footballer's coordination problem (Sugden 2000, 2003).

awareness of one’s status as a team member is neither a sufficient nor a necessary condition of team thinking. It is not sufficient, because one can mistakenly identify oneself as a team member, and it is not necessary, because one can be a member of a team without reflectively identifying oneself as such. In this sense, team thinking is pre-reflective. This also means that if we do correctly think of ourselves as members of some team, this is because we are a team, and not the other way around, as Sugden and those subscribing to a similar view of the role of collectivity in coordination seem to believe.

The second reason is the following. Consider again the driver’s coordination problem. From the viewpoint of the standard theories of team thinking, whether or not “right” is rational ultimately depends on *who one takes oneself to be*. If I take myself to be “one of us” (“the other driver and me”) – that is, a member of a team – right is the rational choice, because right/right is what *we* intend to achieve,and my choosing right is my individual contribution to our shared goal. Whereas for somebody who exclusively optimizes her or his individual desires right is just as hypothetically rational as left. It all depends on one’s identity in terms of one’s reflective understanding of oneself. Identity, one could say, is a matter of self-categorization, and it is prior to rationality.

Thus these theories of team thinking seem to offer a kind of a compromise between our driver’s way of reasoning on the one hand and common sense on the other, in that they hold on to the commonsensical view that it is rational to choose right, while at the same time some tacit assumption that is accepted throughout the debate is jettisoned. Why should it be necessary to prove our driver wrong if one claims that right would have been the rational choice? Sugden, Bacharach, and Janssen seem to maintain the commonsensical claim to rationality without having to bear the burden of proof against our driver. “Right” is the rational choice – but from the perspective of a team member, not from the perspective of an isolated individual. What it is rational to do depends on who one takes oneself to be.[9]

As convincing as this relativizing move might seem with respect to the trouble with the rational fool of our initial example, there are some serious doubts left. Let’s again take the pre-theoretical, commonsensical perspective. If in our everyday understanding we take “right” to be the rational choice, we take it to be the rational choice, full stop. We do not mean something like right is the rational choice “for people who take themselves to form a spontaneous team together with the oncoming drivers,” or “for people who not only individually prefer to avoid a collision, but team-prefer to get by the other.” If we call our driver irrational given all he says about the circumstances of his decision, we do not mean something like “irrational as a team member, but rational from the perspective of an atomistic individual.” Once again, we mean irrational, full stop. Thus it seems that if indeed team thinking

[9] For a clear statement of this, compare Elizabeth Anderson’s “Priority of Identity to Rational Principle”: “what principle of choice it is rational to act on depends on a prior determination of personal identity, of who one is. The validity of the principle of expected utility (maximizing the satisfaction of one’s personal preferences) is conditional on regarding oneself as an isolated individual, not a member of any collective agency” (Anderson 2001: 30).

is involved in coordinative behavior of this kind, and if we are right in holding on to the commonsensical view concerning the rationality of such behavior, a stronger conception of the collectivity of team thinking than the one put forth by Sugden is needed. No matter what our self-image might be, we simply are team members in these situations, and as such we share our intentions and goals. What makes our driver irrational is that he is not aware of who he is: he is not an isolated individual, but one of us.[10] The individualistic social ontology to which these theorists subscribe must be dropped.

In this chapter, I have argued that collective intentionality permeates human interaction down to its very basic modes. It is our capacity for collective intentionality that deconditionalizes rational decision-making where coordination among agents is required. Where there is no incentive for unilateral defection, rational agents will think and act as a team. In order to account for this, however, it is necessary to depart from individualistic assumptions such as those implied in the standard economic standard model of behavior, and to widen the perspective to collective intentionality.

Second, I have argued that a stronger conception than the one to be found in the received literature is necessary in order to develop an adequate account of collective intentionality. I have argued for the first of the following three features of a stronger conception in this chapter, and for the second and third in Chapter 2 above:

1. Collective intentionality is *pre-reflective*. It is not a matter of some reflective attitude of the participating individual (that is, the individual's taking herself to be a member of a team), or "self-categorization."
2. Collective intentionality is *irreducible* to individual intentionality; that is, it does not consist in some set of intentions of the form "I intend. . ." An adequate account is incompatible with *formal* individualism.
3. Collective intentionality is *relational*. An adequate account is incompatible with *subjective* individualism.

What are the consequences of collective intentionality analysis for economic theory? The widening of perspective implied in collective intentionality analysis directly affects the notion of the agent. This meets with other tendencies in "heterodox" economic theory. As John B. Davis has pointed out in his book on the theory of the individual in economics (Davis 2003: 130–149), collective intentionality analysis seems to mesh seamlessly with an increasing unease with the "atomistic" standard model of the agent. In this vein, collective intentionality analysis is particularly attractive because it opens a perspective on social identity and human embeddedness that does not hinge on adventitious stigmata such as birth and destiny. There is a tendency in the received literature to conceive of social identities as fixed entities. As Amartya Sen (2004) has convincingly argued, however, social identities are *made* rather than *discovered*; they are a matter of what we *do*

[10] As argued earlier, this does not mean that we have to be reflectively aware of our "true" nature as social beings in order to avoid collisions in everyday life. Indeed the fact that even individuals who take themselves to be atomistic homines oeconomici can successfully coordinate in real life shows how far our reflective self-image can depart from our pre-reflective way of reasoning.

rather than a matter of *what we are*. For all of the work that needs to be done in this relatively new field of research, collective intentionality analysis seems to be a promising candidate for showing how these identities come about.[11] If the argument developed in this chapter is right, it seems that some very rudimentary forms of “social identity” – that is, shared goals pursued by a team – are at play even in the most transient of our interactions, such as the one of two drivers successfully passing by each other on the highway.

[11] For an interpretation of Sen’s influential criticism of orthodox economic theory in terms of collective intentionality analysis, see Anderson (2001) and Schmid (2005a). For a discussion of the importance of collective intentionality analysis for experimental economics, compare Schmid (2005b).

Chapter 7
Beyond Self-Goal Choice

Rationality and Commitment

Discussions of the cooperation problem, as encountered in experimental economics in Chapter 5, and the comments on the theory of coordination and necessary revisions of our model of practical reasoning in Chapter 6, have shown how important it is to include an understanding of social identity and the sharing of intentional attitudes in social science. This sets us in sharp opposition to the dominant view of economic rationality. At the same time, there are many sources in earlier social theory as well as in the current debate on the economic model of human behavior on which such a revision can draw. In this chapter, one of these sources – perhaps the most important one in terms of personal reputation – shall be examined. In the current debate on economic rationality, Amartya Sen's work plays a uniquely important role. Sen is widely regarded as one of the most astute and thorough critics of rational choice theory; papers such as his *Rational Fools* (1977) have been of tremendous influence on the further development of the debate. In this paper, as well as in his later contributions to the topic, Sen largely relies on one conceptual tool to demonstrate the limitations of rational choice. The concept in question is *commitment*. Commitment, Sen argues, is a central feature of most domains of human behavior. And it cannot be accounted for, Sen claims, within rational choice theory. This chapter examines Sen's claim. Special attention is paid to the way Sen ties commitment to social identity. Moreover, it is argued that the most radical of Sen's claim, which even sympathetic interpreters tend to reject, makes sound sense if we consider the structure of joint action. The issue at stake here is Sen's claim that an adequate account of committed action requires us to go beyond what Sen calls the self-goal choice assumption. This is true in the most straightforward sense, I argue, if we consider the structure of *collective* goals.

I shall proceed as follows. In a first step, I shall present two ways of meeting the challenge set by commitment in the received theory of practical rationality. The first way is the defensive strategy, which sees commitment as an element of the enlarged subjective motivational set of the agent. The second way is to pit commitment against instrumental reasoning. This is the critical strategy, which is chosen by authors such as Robert B. Brandom, and John Searle. I will then turn to Amartya Sen's account of committed action. Sen's thoughts on commitment follow the critical line. Uniquely radical among the claims he makes concerning the relation between rational choice and commitment is that committed action violates the *self-goal choice*

H.B. Schmid, *Plural Action: Essays in Philosophy and Social Science*, Contributions to Phenomenology 58,

assumption implicit in rational choice theory, i.e. the assumption that people should be seen as basically pursuing only *their own goals*. As many of Sen's interpreters have pointed out, this claim seems problematic because it appears that self-goal choice is part and parcel of the folk psychological concept of action. So how could any kind of agency ever violate self-goal choice?

In defending Sen's claim, I shall resort to the theory of collective intentionality. I shall argue that Sen's claim does make sense with regard to *shared* goals. In interpreting Sen's claim, special attention will be paid to the role of social identity in committed action. Committed agents, it is argued, are basically *team players*. This chapter ends with the claim that by construing Sen's concept of committed action in this way, the most obvious problem of other critical accounts of committed agency can be avoided.

§23 Commitment: Two Opposing Views

In the philosophical debate about the limits and scope of rational choice theory, the analysis of the structure of commitment plays a uniquely important role (Weirich 2004: 387ff.). However, Sen is not alone in pitting committed action against the standard model of rational behavior. Before turning to Sen's analysis below in §24, I shall start with an observation concerning some of the other relevant accounts.

It seems that the concept of commitment plays a key role in two opposing views on what is wrong about the classical model. On the first view, commitment epitomizes everything that transcends those egoistic preferences, inclinations, and desires on which Homines oeconomici are usually taken to act. What is needed in order to accommodate committed action is, first of all, a wider concept of the subjective motivational base of actions, and perhaps to allow for a less static conception, which gives more room for deliberation, and for planning (e.g., Verbeek 2007). On this first view, talk about "desires" as being the motivational base of action has to be taken in the sense of Davidson's (1963) "pro-attitudes", or in something like the formal sense in which Bernard Williams uses this term. As Williams puts it, the "subjective motivational set" is not limited to egoistic impulses or desires, but "can contain such things as dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects [. . .] embodying *commitments* of the agent" (Williams 1979: 20; my emphasis). Thus commitment appears as just another form of motivation which, together with appropriate beliefs, rationalizes an agent's behavior. All that is needed to accommodate commitment is a relaxation of the conceptual restrictions on human motivation.

According to the second account, however, "commitment" stands for the necessity of much more radical changes in our understanding of practical reason. On this view, it is not enough to widen our concept of motivation. If commitments are reasons for action, this is not because these commitments somehow express what the agent wants. Commitments are not based in the agent's desires. Quite to the

opposite: if an agent wants what she does when she acts on a commitment, she wants it because she believes she has a reason to do so, and not the other way around. Thus on this second view on committed action, reasons and not motivations are metaphysically basic (cf. McNaughton and Rawling 2004: 117). In this sense, commitment plays a key role in those theories of practical reason which are radically skeptical of the understanding of rationality in action that is usually called "Humean" (even though it might not have as much in common with David Hume's actual views as its proponents like to think). Robert B. Brandom describes his anti-Humean turn in the following words:

> The concepts of desire and preference are [. . .] demoted from their position of privilege [. . .] Endorsement and commitment are at the center of rational agency [. . .] and inclination enters only insofar as rational agents must bring inclination in the train of rational propriety, not the other way around. (Brandom 2000: 30)[1]

Most prominently, John Searle has sketched a non-Humean account of rationality in action, in which an analysis of the structure of commitment plays a key role (Searle 2001a). On his view, commitments do not fit into an account of rationality in action, which bases the reasons for action in the subjective motivations of the agent. Rather, commitments create, as Searle puts it, desire-independent reasons for action. In Searle's example, one does not have to have any (egoistic or altruistic) desire to have reason to pay for the beer one has ordered. The fact that one has ordered the beer is quite reason enough. Searle's analysis of the structure of commitments runs about as follows: commitments are created with the use of language; by means of some "semantical categorical imperative," as Searle calls it, ordering a beer in a bar results in the creation of a reason to pay for the beer, a reason which is independent of whatever the agent in question does or does not have in her or his subjective motivational set (Searle 2001a: 167ff.).

As opposed to the first, Humean or internalist, account of commitment, the second one is the Kantian or externalist one. I do not want to go further into the details of either of these accounts here, but limit myself to the most obvious problems of both views. The problem with the Humean view of commitment seems to be that it blurs the distinction between two different cases of reasons for action. From the agent's point of view, at least, it seems important to distinguish the case in which we believe we have reason to do x because we want to do x from the case in which we want to do x because we believe we have a reason to do x. Sometimes, there are even cases of conflict. One sometimes feels bound by commitments against one's "subjective motivations," however wide these motivations are (one might even feel bound by commitments against one's altruistic motivations). It seems that the Humean view cannot do full justice to these cases.

The existing Kantian or externalist accounts of commitment, in turn, have their own problems. If one takes reasons for action, but not motivation, to be metaphysically basic, especially if one accepts the creation of reasons for action through the

[1] For another non-Humean account of practical rationality based on an analysis of the structure of commitment see Benn and Gaus (1986).

semantic categorical imperative, the old question imposes itself, of how those reasons, all in themselves, should move us to act, without the aid of some desires such as the one to be a rational agent.[2] It is a well-known feature of everyday life that we fail to do what we have reason to do even in cases in which we are aware of those reasons. So what is missing in these cases in which reasons fail to motivate us? In his book on rationality in action, Searle tries to answer this question with what he calls "secondary desires," which are desires, but desires that are created by the recognition of some prior desire-independent reason (Searle 2001a: 168ff.). In other words, those secondary desires play the decisive role of ensuring that one really wants to do what one ought to do.

As such, secondary desires are simply too good to be true. In Searle's story, these secondary desires play the dubious role of the *deus ex machina*, who suddenly puts in an appearance on the scene to save Searle's externalist account. And indeed it seems hard to see why we should worry about the semantic categorical imperative were it not for some prior desire such as the one to be consistent in our views, or the desire to be a trustworthy person and not to erode the base of mutual trust, or some other desire of this type.

Both accounts of committed action have their relative strengths and weaknesses. Perhaps the problem with finding out what side to take has to do with the way the line between the two camps is drawn. Looking at this constellation from afar, I think it is plausible to assume that there might be something wrong with this whole controversy. Maybe the whole question concerning the relation between motivation and commitment is wrongly put. Even though I do not know what Sen's own position on the controversy between internalist and externalist accounts of commitment is,[3] I think that some elements in his analysis of committed action point the way to leaving that constellation behind. In the following, I shall turn to Sen's analysis (§24), before coming back to the controversy between internalist and externalist accounts of commitment at the end of the chapter (§25).

§24 Amartya Sen's Critique of Self-Goal Choice

It seems that in his papers on the topic, Sen's analysis of the structure of committed action revolves around two main ideas, one of which is widely accepted, while the other, as far as I can see, has not met with much approval so far. The first, less controversial point concerns the "wedge between choice and welfare" driven by committed action, which Sen postulates in his paper on "Rational Fools." Committed action requires us to go beyond narrow standard models of preference. "Preferences as rankings have to be replaced by a richer structure involving meta-rankings and

[2] Or, to put it in Amy Peikoff's words: "Rational action entails rational desire" (Peikoff 2003).

[3] In a footnote on the relation between his own "external reference" approach and Williams' internalism, Sen claims to be in line with Williams, because, unlike Williams' internalism, "external reference" externalism is about choice, not about persons (Sen 1995: 30).

related concepts" (Sen 1977: 344). In his paper on "Goals, commitment, and identity," Sen further analyzes this by saying that committed action violates both the assumption that a person's welfare depends only on her or his own consumption (goal-self-regardingness), and the assumption that a person's only goal is to maximize his or her welfare (self-welfare goal), including satisfaction of sympathy. Both assumptions are implicit in the standard economic model of rational action (Sen 1985: 213). Whereas these two points can be seen as a refinement of the earlier statement made in "Rational Fools," Sen now goes one step further by saying that there is yet another standard assumption that is violated by committed action. It is self-goal choice. According to the more radical of Sen's two statements of the self-goal choice assumption (Sen 2002: 34), it basically says the following: "a person's choices must be based entirely on the pursuit of her own goals." (In a slightly softer version, self-goal choice is taken to mean that "each act of choice is guided immediately by the pursuit of one's own goals" [Sen 1985: 214, 1987: 80; my emphasis].) Since, in Sen's view, committed action violates this assumption, the wedge driven by commitment is not between the agent's choice and her or his welfare, as it was in "Rational Fools." Rather, it is between the agent's choice and her or his goals. The claim is that committed agents do not pursue their (own) goals. As Sen knows well, this claim sounds rather extreme. Indeed it seems that in spite of its appeal to some everyday phrases, it is not even understandable. In everyday parlance, we might say of strongly altruistic or heteronomous people that they do not pursue their own goals, but the goals of other people instead. Yet in the proper sense, self-goal choice is not violated even in the most extreme cases. For the whole clue of such strongly altruistic or perhaps heteronomous behavior seems to be that the agent makes the other's goals his own. As Sen, who is well aware of this problem, puts it: "it might appear that if I were to pursue anything other than what I see as my own 'goals', then I am suffering from an illusion; these other things are my goals, contrary to what I might believe" (Sen 2002: 212).

Perhaps the problem in Sen's claim becomes clearer if we take a closer look at the role of goals in agency. I take it that, in a basic sense, goals are something like the conditions of satisfaction of intentions. "Conditions of satisfaction" is meant in Searle's sense (Searle 1983), and it has nothing to do with any kind of psychological enjoyment. The claim that goals are the conditions of satisfaction of intentions simply means the following: goals are whatever has to be the case for somebody to have done what she or he intended to do. In order to attain my goal of closing the door, I simply have of closing the door.

As compared to other, more elaborate accounts of goals and their roles in agency, this approach might seem overly simplistic. More than that, it might appear that this reading draws intentions and goals too close together. Especially, it seems that to identify goals with conditions of satisfaction of intentions wrongfully excludes such cases as when somebody may be said to have a goal without actually intending to do something about it. I might have the goal to close the door, and yet not the intention to close the door, because my more important goal is to eat the ice

cream.[4] Against this objection, one might argue that the intention to do something about it is what distinguishes an actual goal from a mere wish, or desired state of affairs. However, we need not settle this issue here, because in the present context, the role of goals interests us only insofar as goals pertain to intentionality and action (or, in the parlance of the economic model of behavior: to choice). Thus we need not claim that there are no goals without intentions, or no intentions without goals, for that matter (even though I conjecture that the use of the term "goal" in these cases is widely equivocal). All that is claimed is that the role of goals in action is that of conditions of satisfaction of the corresponding intentions. I assume that something similar must be included in any account of the role of goals in agency. And this claim seems especially fit to shed light on the trouble with Sen's critique of self-goal choice.

The example mentioned above may serve to illustrate the point. In order to attain my goal to close the door, I simply have to close the door. This, however, I have to do myself, because the mere fact that the door is shut is not enough to satisfy my intention. If you pre-empt me and close the door for me, or if the draft does the job before I could get around to doing it, this might fully satisfy some other intentional state of mine such as my long-standing desire that the door be closed. However, it does not satisfy my intention to close the door (which might have been prompted by that desire). This well-established fact directly pertains to what is at stake in Sen's claim that self-goal choice is violated in committed action. In a manner of speaking, one can transcend one's own aims in all sorts of ways, for example by intending to do something on behalf of others, or for the benefit of others. Also, one can intend to influence other people so as to prompt them to act according to one's own wishes. However, one cannot directly intend the other's actions, because one can intend only what one takes oneself to be able to do (cf. Baier 1970). I can intend to make it the case that you close the door, but I cannot intend your closing the door (Stoutland 1997). In continental philosophy, this basic feature is sometimes called the "mineness" or "ownness" of intentionality.[5] Just as one cannot die the death of others, even though in some cases, one can die *for* them, one cannot pursue the other's goals without making these goals one's own. This is an essential fact about our intentionality. Thus it seems that what Sen believes to be violated by committed action is nothing less than a basic trait of what makes an agent an agent – at least if we take intentionality as constitutive of agency, and if we take goals to be the conditions of satisfaction of intentions.[6] Or, to put it negatively: no agency without self-goal choice. In this sense, the claim that the structure of committed action (or any action, for that matter) violates self-goal choice seems to be a *contradictio in adjecto*.

---

[4] The example is by courtesy of Peter Vallentyne, to whom I am grateful for pointing out the problem.

[5] "Mineness" translates such terms as Martin Heidegger's "Jemeinigkeit" (Heidegger [1927] 1996).

[6] The last clause is of special importance. Clearly, there is no problem involved in pursuing other people's goals where goals are simply desired states of affairs, rather than conditions of satisfaction of intentions. Concerning the decision for an intentionality related concept of goals, see the above remarks.

Should we therefore simply forget about Sen's second claim, taking it as a condonable excess of his righteous fury at the annoyingly persistent small-minded idea of agency in economic theory? Should we just return to the first feature of Sen's analysis of the structure of committed action, the wedge between choice and welfare, which is less controversial, and still an important contribution to the theory of rationality in action? Or is there any way to make sense of the idea of a violation of self-goal choice by a committed agent?

I suggest that we start by taking a closer look at Sen's claim. In "Rational Fools," Sen already emphasized the role of group membership for committed action. In "Goals, Commitment, and Identity," as well as in other papers, Sen further elaborates this idea. On a first line of thought, Sen introduces "as if" goals to explain the violation of self-goal choice by committed action.[7] However, Sen is well aware that "as if" goals offer no more than a formal equivalent, which does not capture the real structure of the phenomenon.[8] Just the fact that committed action can sometimes be accommodated in "as if" objective functions (Sen 2002: 41), in itself, does not shed light on the structure of committed action. The question is: what do people actually do when their behavior violates self-goal choice?

In addressing this question, Sen introduces the concept of interpersonal or social *identity*. As Sen puts it, "the pursuit of private goals may well be compromised by the consideration of the goals of others in the group with whom the person has some sense of identity" (Sen 2002: 215). It is, as he says, this "sense of identity" which "partly disconnects a person's choice of actions from the pursuit of self-goal" (ibid. 216). One might wonder what this "sense of identity" – which drives a wedge between choice and self-goal – might be. In some passages, Sen seems to suggest a reading according to which the agent identifies himself so thoroughly with another person that the goals he pursues are no longer his own goals. The assumption that one can pursue other people's goals without making them one's own, however, flies in the face of our understanding of agency as analyzed above; taken in this sense, identification amounts to some paradoxical self-elimination. If the object of identification is taken to be some other person, any attempt to go beyond self-goal choice by means of identification amounts to nothing but the futile attempt to stop being oneself by taking on somebody else's identity (cf. Charlie Kaufman's *Being John Malkovich* for a vivid illustration). In this self-eliminative sense, identification with others is simply self-defeating. The harder one tries to get rid of one's own identity by identifying with somebody else, the more it becomes apparent that it is all about *oneself* trying to be another, and not another.

---

[7] "Consider a pair of individuals whose real goals are those as in the Prisoner's Dilemma, but whose actual behavior violates goal-priority (and self-goal choice). The 'revealed preference' relation of their respective choice functions may place the cooperative outcome on top, that is, they may behave 'as if' they would favor that particular outcome most of all" (Sen 2002: 217).

[8] In "Maximization and the Act of Choice," Sen states with regard to the phenomenon of Japanese employees working themselves literally to death: "The as if preference works well enough formally, but the sociology of the phenomenon calls for something more than the establishment of formal equivalences" (Sen 2002: 191).

In this sense, identification is self-defeating, because the very act of identification presupposes the very difference in identity that the agent in question tries to eliminate. On this line, there is no way to go beyond self-goal choice, because no matter how far one goes in making somebody else's goals ones own, it is still invariably *one's own* goals that one pursues.

However, this self-eliminative sense is not the only reading of the role of identification that Sen suggests. The predominant line is quite a different one: here, identification is not with others, taken as single agents. It is not a matter of any I-Thou relation, but between agents and groups – a matter of the I-We relation, as it were. In this sense, identification is not *self-eliminating* (which would be self-defeating). Rather, it is *self-contextualizing*. This kind of identification is not about trying to be somebody else with whom one identifies, but simply about not just being oneself, but one of us. This second concept of identification is the one put forth in Sen's talk on "Reason before Identity", where Sen develops an understanding of belonging that avoids the pitfalls of the communitarian critique of liberalism (Sen 1999; cf. also Sen 2004).

On this second line, the claim that committed action violates self-goal choice takes on a very different meaning. If identification with a group lies at the heart of the structure of commitment, an agent does not have to perform the paradoxical task of choosing someone else's goal without making it his own in order to qualify as truly committed. In a sense, committed action is neither about one's own goals, nor about anybody else's goals. The point seems to be that in committed action, the goals in question are not individual goals, but *shared* goals. If the scandal of the self-goal choice assumption is that it implies too narrow a conception of goals, this is not because it excludes some form of altruism, but because it wrongfully limits goals to individual goals, thereby banning shared goals from the picture. What is needed in order to correct the shortcomings of the self-goal choice assumption is not an account of other-goal choice, but an account of the pursuit of shared goals, or of collective agency.[9] As Sen puts it: "'We' demand things; 'our' actions reflect 'our' concerns; 'we' protest at injustice done to 'us'" (Sen 2002: 215).

§25 Commitment: A Third Account

This "self-contextualizing" notion of identification, however, has its own problems. How does the claim that collective agency violates own-goal choice square with the earlier thesis that self-goal choice is a defining feature of any kind of agency? If the

[9] For an analysis of the link between Sen's concept of identification and the demand for a robust concept of collective agency, see Anderson (2001). In her reflections on collective agency, Carol Rovane clearly distinguishes projection into another individuals' points of view from orientation on common ends: "these activities do not require that persons project themselves all the way into another person's own rational point of view so as to take up that person's perspective. These activities require rather that persons project themselves into a rational space that is generated by the ends which they hold in common [. . .] When persons project themselves into this common rational space, they can reason and act together from the perspective of their common ends" (Rovane 1998: 138).

earlier considerations on the status of goals in intentional behavior are correct, it seems that departing from self-goal choice amounts to endorsing one of the following two equally repellent alternatives. Either it requires denying that the individuals taking part in collective actions are proper agents, or it requires making a category mistake of the most basic Rylean type. The first of these alternatives seems implausible because whatever one takes collective action to be, it is clear that the individuals involved in shared activity are agents, not just, say, organs in some collective body. There is no reason to doubt that it is legitimate to demand that an account of collective agency be consistent with the notion that individuals do act when they act together. If one accepts this assumption, however, it appears that the only reason left to believe that collective agency violates self-goal choice is a category mistake. For the only alternative then seems to be to understand collective action as something different from the actions of the participating individuals. This, however, is in direct conflict with the predominant view, according to which it is not only the case that individuals act when they act together, but that the actions of the participating individuals is what collective agency is. There is no collective agent, no macro-subject, that acts in addition to the participating individuals, when individuals act jointly. To adapt the Rylean example to the given case, it seems that whoever contests this makes a mistake similar to the spectator watching some soccer game for 90 min, before saying "I have had enough now of those twenty-two people running about on the field in some coordinated way. I just wonder when, finally, the teams will start playing!" Because individuals, running about on the field in some coordinated way is what team play *is*.

Therefore, it appears that collective agency does not violate self-goal choice: all that is chosen in collective action is individual goals, namely the goal to contribute to the attainment of some shared aim. As it was put in an earlier contribution to the theory of shared goals: if a team has goal x, than each individual member has goal x (cf. Levesque and Cohen 1991) – or, more precisely, some contributive goal y – which conforms to self-goal choice.

Thus it seems that any attempt to depart from self-goal choice faces a dilemma. It amounts to ending up either in some massively collectivist conception, which flies in the face of even our most basic understanding of intentional autonomy (cf. Pettit 1996: 117ff.), or in a conception that is based on a simple category mistake. Since both alternatives appear equally unacceptable, it seems that we should not depart from self-goal choice.

I think, however, that the argument concerning the second alternative is not sound. In the following, I shall argue that even though the participants act when they act jointly, there is no category mistake in assuming that joint action violates self-goal choice. The thesis I would like to put forth is not that agents violate self-goal choice when they act together (this claim would lead directly into some of the nonsense we have encountered before). Rather, my claim is that the self-goals which individuals choose when they act together cannot be adequately represented within an account which takes all goals to be self-goals, because these self-goals presuppose shared goals.

The argument is the one put forth by those advocating a non-reductivist reading of collective agency. Participative intentions and goals are, to use a term coined by Wilfrid Sellars, "we-derivative" (Sellars 1980: 99). If we play a duet together, my aim is not just to play my part while you play yours (such cases may occur, but they do not constitute genuine cases of shared agency). Instead, it is as a part of our shared activity that you and I do what we do individually when we play together (cf. Searle 1990). In order to account for our contributive self-goal choices, an observer needs to understand that what she or he observes is something the agents are doing together (for more arguments for the non-reductivist view cf. Chapter 2 above).

Some current accounts of shared agency and collective intentionality are accused of circularity, because their analysis of what individuals do when they act together presupposes what should be explained. From a non-reductivist perspective such as the one I just have taken, this is not surprising, but simply reflects the ontological structure of participative intentions or participative goals. In the sense of the "we-derivativeness" of participatory intentions and goals, togetherness is irreducible; or, to use Sen's term of the "privateness" of goals: shared goals are not simply combinations of private goals. There is a difference between goals that individuals just somehow happen to have in common, on the one hand, and goals which individuals have individually only because they have this goal in common, on the other.[10] An account of agency that is unable to see beyond the limits of self-goal choice cannot account for the latter kind of goals, i.e., the case of genuinely shared agency. Paradoxically, the self-goal choice assumption renders action theory blind for one special, but important kind of self-goal choice, namely, contributive self-goal choice.

There is yet another argument for a non-reductivist account of collective agency that I would like to mention, even though this brings me into some tension with Raimo Tuomela's account of collective agency. As Annette Baier (1997a: 26, 1997b: 37) has pointed out, there are some rare cases in which individuals fail to form an appropriate we-derivative individual intention, even though, in a sense, they still can be said to share an intention (for a differing view cf. Tuomela 1991: 271ff., 1995: 135ff.). Take the case of some spontaneous and transitory collective action, such as the one of a couple of passers-by joining their forces in order to push a car. As a participant in that activity, I might suddenly feel estranged from my role and lack the aim to provide my contribution, even though I might still think of our goal to push the car as our goal, and not merely as their, the other people's, goal. In such cases, it seems to make perfect sense to speak of collective goals or collective intentions in a sense that does not refer to corresponding individual contributive goals or intentions. An account that is based on self-goal choice seems to be blind for such cases.

[10] Jay Rosenberg calls the former type of ends "common" and the latter "communal." "A communal end. . . will be one which is collective without being conjunctive. It will be an end which is mine and hers and his by virtue of the fact that it is ours and that each of us represents himself/herself as one of us. It will, in other words, be a genuinely plural end, attributable to all of us collectively and therefore univocally to each of us severally and to all of us conjunctively" (Rosenberg 1980: 160).

Admittedly, these are rare and perhaps even pathological cases. But in light of such deviant cases, normality reveals some of its basic traits. If I think of some goal as our goal, I can be expected to have a corresponding individual contributive goal, or some other kind of pro-attitude. In the absence of overriding reasons, I should choose to do my part. The relation between shared goals and individual contributive goals (i.e., between shared goals and self-goal choice) is a normative one. This, however, points against a constitutive relation between individual contributions and shared goals of the kind at work in reductivist accounts of collective agency. Normativity entails contingency. That I should choose my contributive goal in our collective project presupposes the possibility that I decide not to contribute to the attainment of our shared goal. The possibility (perhaps more than the fact) of dissidence, as well as of other kinds of failures to do one's part, is an essential part of shared agency. It is what makes the relation between shared goals and individual choices normative. And again, an account that is limited to self-goal choice seems to be blind to the fact that some self-goal choices normatively depend on shared goals. In short, the self-goal choice assumption is incompatible with a nonreductivist account of collective agency.[11]

As was pointed out early on in the collective intentionality debate, shared intentions or projects provide us with a standpoint from which we critically measure and evaluate our individual plans and aims (Rosenberg 1980: 159). As normative sources, shared intentions, aims, goals, and projects provide us with reasons for individual action. This brings me back to the initial point concerning the controversy between internalist and externalist accounts of commitment. For these special reasons, which are based in shared intentions and projects (in short: shared desires), have an interesting status. They are neither internal nor external reasons. In some sense, they are independent of us as single individuals, or, more precisely, they transcend our "subjective motivational set" – that is why they can serve us as a critical standpoint for our self-evaluation. In this sense, reasons that accrue from shared desires are not internal. On the other hand, these reasons are not external either. They are not disconnected from the sphere of "desires" (in Williams's formal sense of the word). If and insofar as the reasons for committed action are ultimately based in shared desires, the distinction between internal and external reasons does not apply. Because shared desires are neither internal to one's motivational set, nor external. Instead, they transcend one's subjective motivational set. An account of the structure of commitment that has neither "subjective motivations" nor "metaphysically basic" reasons, but shared desires playing the leading part in committed action, seems to avoid the two problems I have mentioned at the beginning of this chapter. It avoids both the "Humean" inability to conceive of the agent's power to transcend their individual desires, and the old "Kantian" problem of first throwing motivation out with some great gesture of depreciation and then having to beg it in again through the back door.

[11] Thus I assume that the self-goal choice assumption is ultimately equivalent to what Margaret Gilbert (1989: 418–425) criticizes under the label "singularism".

In the rich literature on Williams' internalism about practical reason, it seems that Martin Hollis' view is closest to the one developed here. In spite of his externalist bias, Hollis comes close to an account of shared desires, when he discusses the relation between "interest" and community (Hollis 1987). If we move from interest to shared desire, the problem with Williams' internalism is not that it bases reasons in motivation. Instead, it is the way in which Williams conceives of human motivation. Not all our motives are part of our "subjective motivational set." Some are intersubjective. I believe that this insight is part of what makes Sen's invitation to look beyond the limits of self-goal choice so important.[12]

[12] I am grateful to the participants of the Workshop on Rationality and Commitment, held at the University of St Gallen on May 13–15, 2004 (especially to Raimo Tuomela, and Philip Pettit), to Peter Vallentyne, and to the two referees for Oxford University Press for their criticism and comments on this paper.

Chapter 8
Lending a Hand

The Structure of Everyday Cooperation

In conclusion of the second part of this book, I shall come back to a topic that has been lurking in the background of all previous chapters of this section: the question of altruism. I have argued above that the concept of altruism is insufficient for capturing the motivational structure of cooperation (Chapter 5), coordination (Chapter 6) and commitment (Chapter 7). It is now time to address the question concerning the structure of altruism. Over the past decade, the concept of altruism has come to play an increasingly important role in social science. This is particularly true in experimental economics, where altruism is routinely quoted when it comes to explaining the vast discrepancies between the observed behavior in the experiments, and the predictions based on the standard economic model of human behavior. In this debate, as well as in some other contexts, altruism usually means having 'pro-social' or 'other-directed' preferences. It is observed that people are not always egoistic in terms of the somewhat narrow conception of self-interest that is still at work in much of economic theory. The conclusion that is usually drawn is to drop the assumption that individuals are only interested in what they can get for themselves in favor of a wider conception that extends to such preferences as benevolent desires, preferences for reciprocity and fair dealing, and the inclination to punish transgressions against the norms of fairness even if one is not directly affected by that transgression, and if punishment is costly to the punisher (cf., e.g., Fehr and Fischbacher 2003; Henrich et al. 2004).

Thus the term 'altruism' has become something of an indicator for what one might call the *defensive strategy* in recent economic theory. Proponents of this strategy acknowledge systematic deviations between the economic model and actual human behavior, but tend to believe that it is possible to correct this shortcoming simply by widening the class of human preferences. Not all participants in the debate, however, believe that such amendments to the model will do. Authors such as Amartya K. Sen have voiced serious doubts concerning this defensive strategy (Sen 1977, [1985] 2002; Peter and Schmid [eds] 2007). These authors favor what one might label the *critical strategy*, claiming that much more radical conceptual changes than a simple expansion of our view of human desires will be needed in order to come to an adequate understanding of human action, changes that affect not only our notion of motivation, but our concepts of the agent's identity and the nature of choice, too.

H.B. Schmid, *Plural Action: Essays in Philosophy and Social Science*, Contributions to Phenomenology 58,

Put very bluntly, the general thrust of this paper is to reclaim the term 'altruism' for the critical camp. The main claim is that limiting the extension of the term to such phenomena as benevolent desires or other-regarding preferences, as it is done within the defensive venture, means pulling the concept's teeth. An adequate theory of altruism has to go beyond pro-social and other-directed preferences, and necessitates far-reaching revisions in our outlook on human behavior.

The concept of altruism is approached from the perspective of action theory. The point of departure (in §26) is what I call the paradox of altruistic action, i.e. a conflict between the concept of action as used in much of current social science on the one hand, and our intuitive notion of altruism on the other. It seems that some of our pre-theoretic intuitions concerning the structure of altruism do not square with our standard theory of action. In the current literature, it is usually believed that this paradox is due to an overstrained notion of altruism, and that a more relaxed view of altruism can easily be accommodated in the standard theory of action via other-directed desires. The next section voices doubts concerning this solution to the paradox. A class of behavior is discussed which does indeed seem to be altruistic in the strong intuitive sense, and does not seem to fit other-directed desire explanations. The distinguishing feature of this class of altruistic behavior is the peculiar way in which the benefactor's behavior is linked to the beneficiary's pro-attitudes. (In the following, the term 'benefactor' refers to the altruist, 'beneficiary' stands for the individual or group of individuals profiting from the action in question.) The paradigmatic case for this class of behavior is the case of spontaneous, low-cost and transitory supportive behavior among strangers, such as moving aside to facilitate another person's passage, or holding a door for another person in a railway station. In everyday folk psychological explanations, we tend to explain such behavior in terms of the beneficiary's pro-attitudes rather than in terms of any of the benefactor's own. I call such explanations heterodox. Heterodox explanations go against the grain of a basic action theoretic assumption according to which behavior always has to be explained in terms of the agent's own pro-attitudes. This classic view is expressed most clearly in Donald Davidson's action theory (the *locus classicus* being Davidson 1963), and it is a basic feature of the "Humean" model of action, and there seems to be no version of intentional and rational choice explanations that does not rest on this assumption (cf., e.g., Elster 1985). If heterodox explanations of cooperative everyday behavior are literally true, much of received action theory is wrong. At the same time, the heterodox view is not limited to folk psychology and everyday talk. It seems to receive some support from two more sides at least. On the one hand, it is in tune with some old and venerable strands in the theory of empathy (Lipps 1903). On the other hand, there are some psychological theories – one regarding the link between empathy and altruism (Batson 1994), the other one regarding cooperative behavior in early childhood (Tomasello 1998) – that seem to lend some support to heterodox explanations of the behavior in question (or so I shall argue). Therefore I shall conclude that we should treat the interpretation of the behavior in question as an open issue, and grant heterodox explanations the benefit of doubt.

The final section comes back to the paradox of altruism, and addresses the hypothetical question: *if* the heterodox explanation is correct, could the behavior in question still be interpreted as the benefactor's own action? Or would the benefactor's behavior have to be attributed to the beneficiary's agency, just as metaphors such as the colloquial expression "lending a hand" seem to suggest? I argue that heterodox explanations need not displace the benefactor's own agency. The argument uses an important distinction, which could be useful for further refinement in action theory even if heterodox explanations should turn out to be wrong. The distinction is between individual intentional autonomy – i.e. the claim that each individual's behavior instantiates his or her own action – and individual motivational autarky, i.e. the claim that the explanation of each individual's behavior has to bottom out in that individual's own pro-attitudes. I argue that to reject the latter assumption does not entail rejecting the former.

§26 The Paradox of Altruistic Action

Let me start by mentioning some basic features of the concept of action that I take to be fairly uncontroversial, or at least acceptable for most analytical philosophers, social scientists, as well as for most competent speakers of ordinary language. First, for there to be action there has to be some kind of *agent*. The role of the agent is at least threefold: in most paradigmatic cases of action, the agent functions as a *source of initiative*; in all cases, he exerts some degree of *control* over the action, and functions as the entity to whom the action is *attributed*, and who can be held responsible for its consequences according to the set of social norms that constitutes or regulates the practices in which the agent participates. Typically – but not necessarily – the agent is an individual. Second, there has to be something that the agent initiates and controls, and for which she can be held responsible. Action requires some kind of *behavior*, which is basically some kind of event or series of events in the world of which the agent is in a degree of control. In the case of external action, the basic events are bodily movements which the agent does not perform by doing something else. The control need not be total; people can act even with trembling hands, just as long as there is some control involved. Third, some *goal* is needed. There has to be something the agent's behavior is all *about.* Here, some qualification is needed. As I use the term in the following, goals are not simply states of affairs the agent *wishes*, or *wants*, or *desires* to exist, or has any other kind of pro-attitude about. Rather, they are whatever has to be the case for somebody to *have done* what he or she *intended* to do. Thus goals are the conditions of satisfaction of intentions, i.e. states of affairs *as being caused* by an intention (Searle 1983). The difference between a desired state of affairs that is not a goal and a goal is this: if I do not simply desire that the door be closed, but *intend* to close the door, and you pre-empt me and close it for me, I do not achieve my goal, even though the state of the world which I desire is

now realized.[1] Thus it seems that goals are a subset of the class of desired states of affairs: agents need to *want* to do what they intend to do, but it is not the case that each of their wishes or desires gives rise, or motivates, a corresponding intention. Agents can have a desire without intending to do anything about it.

Thus desires play a dual role regarding intention. First, desires *constitute* intentions in that they capture the motivational element that is an essential part of intentions. If we assume a *narrow* concept of desire, it is true that one can intend to do things which one does not *want* or *desire* to do, such as in the case of one's intention to keep one's annual dentist's appointment (Schueler 1995; Searle 2001a). But in the current literature, "desires" extend beyond the class of objects the thought of which tends to induce a positive affective reaction. Most of the current literature uses the term in a wide sense that comprises pro-attitudes of any kind: wishes, interests, projects, commitments, inclinations, and so on (Davidson 1963: 685ff.).

The second role which desires play with regard to intentions is *motivational* rather than constitutive. The intention to see the dentist, which encompasses the (extrinsic) desire to do so, is *motivated* by one's intrinsic desire not to suffer the pain that will result from carious teeth. In this second, *motivational* rather than *constitutive* sense, desires logically *precede* intentions and provide the *motivating reason* for forming an intention which is quoted in the explanation of an intention. It is usually assumed that, in order for there to be an action, a linguistically competent agent has to be able to come up with an answer to the question of *why* she or he wanted to do what she did, and that this answer has to quote some volitional agenda of her own.[2] So, in one sense, desires describe the constitutive motivational *component* of intention (i.e. the fact that intention is a motivation-encompassing attitude), and in another sense, desires describe some necessary *antecedents* of intention, i.e. the motivational base that explain the intention.

For a desire to be able to explain an intention, however, yet another feature has to be in place. The intention (and therefore the complex of events which it causally controls) has to be *minimally rational*, i.e. the agent has to show at least some minimal degree of concern about the intention qua executive plan being conducive to the end (however successful or unsuccessful she might be at this task). To put it somewhat more cautiously: the agent cannot be *entirely indifferent* as to whether or not the induced events are suitable as a means to achieving her goals, and as to whether or not there are better means available to her. She has to *believe* that her behavior is a suitable means to realize her goals. To put it in Davidsonian terms, the *primary reason* that rationalizes the agent's behavior has to include some pro-attitude, and some suitable belief.

---

[1] As is obvious from this terminology, the concept of intention used in this chapter is largely Searlean (Searle 1983). As far as I can see, however, nothing of what I say is in conflict with a view of intention in terms of executive plans along more Bratmanian lines (c.f., e.g., Bratman 1999).

[2] It seems that the more the desire that *constitutes* the intention is *intrinsic* and *general*, the more difficult it becomes to distinguish it from the *motivating* desire that explains the intention: it is difficult to come up with anything else in explanation of one's wanting to lead a meaningful life than just that. But even if there are descriptions under which the distinction collapses, it remains meaningful; in those cases, the constitutive desire *is* the motivating desire.

These four features can all be drawn together in a single phrase. For there to be an action, there has to be some agent-controlled complex of events of which it is possible to make some (minimal) sense in terms of what the agent *wants*.[3]

If this is action, what is the problem with its being altruistic? Why should there be a paradox in the notion of altruistic action? The problem that gives rise to the paradox is this: there seems to be a basic element of *selfishness* built into the very concept of action. And this element, it seems, is at odds with an intuitive idea of genuine altruism, so that a complex of behavior cannot at the same time instantiate an action and a case of genuine altruism. To illustrate this paradox, I borrow from Thomas Nagel's analysis of altruism (1970: 80–81) in the following.

First, the element of selfishness: according to the standard theory, we cannot make sense of a complex of behavior in terms of anything other than whatever the agent happens to want *him-* or *herself*, if we are to interpret that complex of behavior as an action. In other words: behind every action are the agent's *own* desires (in the wide sense of the term explained above). This, however, does not square with an intuitive notion of genuine altruism. According to this notion, genuine altruism should be about the *beneficiary's* interests rather than about any of the benefactor's own, so that the interpretation of *altruistic* behavior should appeal to *other people's interests* rather than to the agent's. This is not altogether implausible: what makes an individual an altruist is precisely that it is possible to make sense of a more or less substantial part of her behavior in terms of the interests of people *other* than herself. And this is precisely how genuine altruists tend to explain their behavior: they did what they did *because other people wanted or needed it to be done*, full stop.[4] Why should we add some of the altruist's own interests to the story, if they don't appear in the altruist's own account?

It might seem that we moved too quickly from pro-attitudes to interests, though. Not all of our pro-attitudes are in our interest, and not all interests are reflected in our pro-attitudes: after all our wanting to breaking our bad habits might be in our interest without us *wanting* to break them, or having some other pro-attitude towards breaking them. And paternalistic cases of altruistic behavior shed a rather sharp light on that fact. In contrast to desires, interests involve the problem of *justification* (which shall be addressed below). But let us disregard for the moment cases in which altruists further their beneficiary's interests against their beneficiary's pro-attitudes. Let us concentrate on those cases in which desires do not collide with interests, and in which it is in the beneficiaries' interest to have their desires fulfilled.

With this in mind, we can now state the paradox. It is a conflict between two propositions that seem plausible at first sight, and from which we seem to be forced

[3] Needless to say, these conditions are *necessary* rather than *sufficient* for the standard concept of action. We do not, however, need to delve deeper into the analysis here as these conditions alone give rise to what I shall call the Paradox of Altruistic Action.

[4] The French phenomenologist Emmanuel Lévinas is famous for making this idea the point of departure of his thinking on interaction and society (cf., e.g., Lévinas 1991). For a lively description of the *immediacy* and of the *unthinking character* of altruistic responses see Craig Taylor's analysis of the structure of sympathy (Taylor 2003). I shall address the objection that such explanations quote *justifying reasons* rather than *motivating reasons* below.

to conclude that there is no such thing as an altruistic action. The two propositions are the following:

(i) A necessary condition for a complex of events to be an action is that, at the basic level of intentional explanation, it can be made sense of in terms of the agent's own pro-attitudes and beliefs.
(ii) If there is no conflict between pro-attitude and interest, a complex of behavior is genuinely altruistic to the degree that, at the basic level of explanation, it is to be made sense of in terms of the beneficiary's desires rather than in terms of any of the altruist's own.

Proposition (i) captures the "Humean" model of action, while proposition (ii) captures an intuitive notion of genuine altruism. To the degree to which we grant both propositions some plausibility, we might see ourselves pushed towards the conclusion that the very notion of altruistic action is an oxymoron. It seems that from a standard action theoretic perspective, a genuine altruist's behavior simply cannot instantiate the altruist's own action: his behavior would have to be attributed to the other's agency rather than to her own, since it is in terms of the other's 'desires' rather than her own that sense can be made of her behavior. If the altruist's behavior is to be taken to instantiate the beneficiary's action rather than the altruist's, it isn't altruistic, since it is rationalized by the agent's own pro-attitudes. Genuine altruism and agency are, it seems, conceptually incompatible. Either an individual is a genuine altruist, or she is an agent.

This paradox is clearly of the Zenonian kind. Since cases of altruistic action abound in real life, there *must* be something wrong with this way of putting things. The question then is: which one of the two conflicting sides is at fault? Is the theory of action with its element of selfishness to blame, or is the intuitive notion of altruism simply skewed? Should we relax proposition (i) or rather proposition (ii)?

Where this question is addressed at all in the current literature, the recommendation is unanimous: proposition (ii) is at fault. It is believed that the apparent conflict is due to an overstrained notion of altruism, which should not be taken seriously. Here are two important examples for this view. Eliot Sober and David Sloan Wilson (1998: 223) claim that to define selfishness in terms of "whatever people want" is simply short-circuited and grossly biased, leading, as they put it, to an utterly "spurious" view of altruism, because this definition leaves no room for motifs that are the agent's own, but are not aimed at the agent's own welfare. Philip Kitcher's remarks on the matter are even harsher. According to Kitcher, the whole paradox is something of a scam anyway, and only non-philosophers could ever be so naïve as to think that there is more to the problem than simple conceptual confusion (Kitcher 1998: 291). According to the line followed by Kitcher and a great many other authors, it is a mistake to think that, just because an interpretation of altruistic behavior has to appeal to other people's interests, it cannot be interpreted in terms of the altruist's own desires. There is, as they point out, such a thing as *altruistic* (or other-directed) *desires.* The view that is often claimed to be commonsensical and even folk psychological is this: altruists are normal agents, just that they have nicer desires. As with any other agents, they do *what they want*, but what they want

happens to be to further other people's interests. Thus there seems to be a way to accommodate the central feature of the intuitive notion of altruism within standard theory of action. Other people's desires *do* guide (and thus to some degree explain) the altruist's behavior, but they do not thereby *displace* the altruist's *own agency*. Rather, the link between the altruist's own behavior and the other's desire is precisely what the altruist *wants.* This link is the condition of satisfaction of her altruistic intention. Thus the respective behavior can be altruistic, and instantiate a case of the altruist's own action at the same time. According to this view, there is no paradox of altruistic action.[5] All that is needed is a clear understanding of the concept of other-directed desires. Other-directed desires are a particular kind of second-order desires whose content is the promotion of other people's interests.

§27 The Structure of Everyday Altruism

I do not dispute the existence and importance of other-directed desires. The claim I wish to defend here is the following. While the paradox of altruistic action can be resolved by relaxing proposition (ii), i.e. by appeal to other-directed desires in a *wide range* of cases, there is another class of apparently altruistic behavior, which cannot easily be fitted into this view. For this class of behavior, we have to find another solution to the paradox, or so I shall argue.

The class of behavior I have in mind here is different from action based on other-directed desires in the following respect. Whereas actions based on other-directed desires are relatively *complex,* requiring second-order desires, some degree of deliberation, and a clear understanding of other people's pro-attitudes, the behavior in question here is very simple, it is unthinking, linked to other people's immediate *goals* rather than to their desires, and in many cases it might appear more like mere reflex behavior than like a proper choice of a course of behavior.

Let me make a short detour to introduce the phenomenon. The idea for this chapter goes back to a session of the Economic Science Association at the ASSA-meeting in Chicago early in 2006. The economist and behavioral scientist Herbert Gintis opened his talk with a simple case of everyday behavior. If I recall correctly, the anecdote went something like this. When he couldn't find out how to open the door to the conference building, some passer-by who observed the scene took it upon herself to quickly press the open-door button for him, immediately leaving the scene after having helped without even waiting to be thanked. I certainly do not want to underestimate either the psychological costs of the sight of Herbert Gintis' plight, or the rewards of his grateful smile, but it seems plausible that such behavior is indeed both genuinely altruistic, i.e. not motivated by any psychological reward

[5] For the purposes of this chapter, I label this solution the *other-directed desires explanation* of altruistic behavior; I will call "other-directed" such desires (in the wide sense) as the desire to help concrete others, or groups of others, as well as such desires as the desire to conform to social norms or rules of conduct.

or cost, and quite pervasive in social life. On my way back home from the meeting, I observed cases such as the following: People holding doors for strangers carrying suitcases; people helping strangers move baby carriages in and out of trains; people moving aside on their benches so that other people could have a seat, too; people facilitating other people's passage by moving out of their way; people lifting other passenger's suitcases to and from carry-on luggage trays; a person picking up an umbrella that had slipped out of the elderly owner's hand; and, as a last example, I overheard a person trying to finish another person's sentence who got stuck in the middle of her question for directions.

I assume that such behavior permeates our entire social lives, but it is in the public sphere that its genuinely altruistic nature becomes most obvious. The observed interactions occurred among complete strangers, and typically, people didn't even wait to be thanked, but simply moved on immediately after having helped. There's genuine altruism at work here, it seems, and not just a hunt for grateful smiles.

Let's first have a closer look at how such behavior is *different* from those cases of altruism which are in the focus of much of the current literature on the topic. Compare such behavior with the paradigmatic case of altruism, which is donating to charity. At least three distinctive features immediately hit the eye. First: while it is essential to paradigmatic cases of altruism that the benefactor incur some costs (of whatever nature they might be), this does not seem to be central in the case of the behavior in question. Indeed, some degree of *indifference* seems to be the hallmark of the observed behavior. The costs incurred by the altruist are minimal, and they certainly play no role in the benefactor's own perception of the situation. This is very different from the donor's case, where some degree of self-sacrifice is part and parcel of the matter.[6] Second, something like *non-deliberativeness:* Paradigmatic acts of altruism typically involve some *care* or *concern* for the people being supported.[7] This entails that, in these cases, the benefactors are *conscious* of the beneficiary's needs, often going as far as to develop an understanding of the beneficiary's needs that differs from the beneficiary's own perception thereof (leading to vicarious, patronizing or patriarchal forms of altruism). This is very different from most of the cases listed above. There doesn't even appear to be enough *thinking* involved for any talk of care or concern to make proper sense. Many of these acts resemble unthinking, spontaneous, perhaps even impulsive behavior much more than deliberate choices. And third, the main difference: *other-goal orientation*. The behavior in question is not directed towards any of the other people's deeper needs, or well-being. Rather, it is just about other people's immediate *goals*. The benefactors in the above example support the beneficiary's in their pursuit of their immediate aims, independent of any evaluation of these goals. By contrast to paradigmatic cases of altruism, the behavior in question is more a matter of manners than a matter of morals. In short, the phenomenon is this: altruistic, low-cost, more or less spontaneous and non-deliberate behavior in pursuit of other people's goals.

[6] The people in the examples I mentioned before wouldn't risk missing their flight; serious cases of people in need of help are left to professional altruists.

[7] Often, this is even made an element of the very definition of the term "altruism".

In a next step, I use the contrast between paradigmatic cases of altruism and Gintis-class behavior to try to raise some doubts concerning the other-directed desire model's explanatory capacity. I have to admit from the outset, however, that I do not have conclusive evidence that this behavior does not fit the model. All I have are two reasons for doubt.

The first clue draws on everyday intuitions, and on ordinary language. I call it the argument from folk psychology. We usually talk about paradigmatic cases of altruism differently than we talk about the kind of behavior in question here. When social psychologists asked people who donated to charity or did volunteer work why they did so, they answered that they "wanted to do something useful" or that they "wanted to do good deeds for others", or something along these lines (Reddy 1980, quoted in Sober and Wilson 1998: 252). These self-reports are perfectly in tune with other-directed desires explanations, quoting those altruistic goals which the altruists themselves wanted to achieve. I do not know if any such study has ever been carried out, but I think it is not implausible to assume that if asked a similar question, a substantial number of the helpers in our cases of everyday altruism would have given answers of quite a different type. If asked why she pushed the open-door button, Herbert Gintis' helper might have replied something like "because (I saw that) he couldn't find it." It seems that this would have been much more natural a reply than something along the lines of "because I wanted to help him (pass the door/enter the building)". Similarly, if one asked the person on the park bench why she had moved aside a little, she would probably say "because he wanted to sit down, too" rather than "because I wanted that he could sit down beside me", "because I wanted to be nice to him", "because I wanted to avoid a conflict with him", or anything of that sort. The decisive difference is this: in explaining the behavior in question, these reports quote *other people's pro-attitudes* rather than any of the agent's own. As opposed to donators, everyday altruists are more likely to explain their behavior in terms of what *other* people want rather than in terms of any of their own other-directed desires. In other words: it seems somewhat artificial to fit this behavior into standard action theory by postulating an additional set of desires which the benefactors do not seem to know of themselves.[8] And if this is true, it seems that what we have here is a case of genuine altruism in the strong intuitive sense mentioned above.

The question, of course, is: how literally should we take such manners of speaking? Self-reports and ordinary language are not the ultimate source of authority in action theory, let alone mere conjectures concerning *possible* self-reports. But it remains a remarkable fact that, in everyday life, people quite often *do* refer to other people's pro-attitudes rather than to any of their own when they are asked to explain their behavior. Especially with Ockham's razor in mind, the question should be asked: what reasons do we have to assume more desires to explain these people's behavior than those they quote themselves when they give an account of what they do? What reasons, that is, besides the fact that this happens to be what standard action theory requires us to do?

[8] It is true that not all desires need be conscious. But it is plausible to assume that desires cannot be *inaccessible to consciousness* (cf. Searle 1981), so that under suitable circumstances, agents are conscious of their desires.

Yet there are at least two obvious objections to the line of argument developed so far. The first objection is that the behavior in question should be interpreted in terms of some desire that one's behavior be guided by the rules of politeness, or some such norm-oriented motivational states from the altruist's part. The person moving aside on the park bench might not have the desire to have the other person sit beside her. But surely, it might seem, she will have the desire to be polite. The reason why I'm not convinced by this objection is the following: whereas most forms of the type of behavior to which our cases of everyday altruism belong are sustained by social norms of politeness and propriety, we can easily find cases in which such behavior actually *violates* social norms. I shall come back to this below.

The second objection is more fundamental. It is this. It might seem that the whole problem with the paradox of altruism arises from an equivocal use of the term "explanation" in the exposition of the paradox of altruistic action above. Whereas in proposition (i) above, "explanation" refers to *motivating reasons* of the behavior in question, it is about *justifying reasons* in proposition (ii). Justifying reasons differ from motivating reasons in the following respect. Whereas motivating reasons *rationalize* a given complex of behavior, *justifying reasons* are facts that seem to make a given goal worth pursuing (cf., e.g., Pettit and Smith 2004: 270). It seems plausible to assume that the two kinds or reasons are mutually independent, and that they may be different kinds of entities. Whereas beliefs and pro-attitudes seem to be the only plausible candidate for *motivating* reasons, it might appear that *all sorts of facts* may act as justifying reasons. Thus it seems that when explaining their behavior in terms of the needs or pro-attitudes of *other people*, altruists refer to the *justifying reasons* for their actions rather to their motivating reasons; they answer the question of why it was *right* or *morally required* to act the way they did rather than the question of what kind of pro-attitude rationalizes their behavior.

In order to assess the validity of this objection, it seems useful to modify some of the above examples so that it can be excluded that the "why"-question is answered in terms of justification rather than motivation. The following thought experiment was suggested to me.[9] Take again the case of Herbert Gintis standing in front of the closed door with his helper observing the scene. But now suppose that there is another person on the scene, the helper's colleague, who is much closer to the button and whom the helper knows to be familiar with the opening mechanism, the helper sees that his colleague observes that Gintis cannot find the button. But the colleague remains inactive, so the helper steps in and pushes the button.

In this situation, the question "why did you push the button" acquires a different meaning (Garfinkel 1981: chap. 1): the question is not "why did you push the button *rather than doing nothing*", but "why did *you rather than he* push the button". As both the helper and her inactive colleague seem to have the same *justifying reason* for action, it seems clear that the explanation for the difference between the helper's and her colleague's response has to be given in *motivational* rather than justificatory terms.

[9] I wish to thank an anonymous referee for Economics and Philosophy for this suggestion.

The question now is: what would the helper's response be? Would she say something along the lines of "Because I want to be helpful to others when I can, whereas he (the non-helping colleague) lacks any such desire", thereby quoting one of her own pro-attitudes? I'm not convinced at all. If the question is simply "why did you push the button" in the presence of the colleague, the most natural reply seems to be "because he (Gintis) wanted to enter the building and my colleague didn't help him". If the question is "why did you *rather than he* push the button", the most plausible reply seems to be something along the lines of "because he (the colleague) was inattentive/wasn't paying heed/didn't bother". None of these answers quote any of the helper's own pro-attitudes.[10]

I conclude that ordinary language and folk psychology does not support the claim that where justifying reasons are not the issue, explanations can only be given in terms of the agent's own pro-attitudes. And I argue that pro-attitudes are not needed to explain the difference between everyday altruists and apparent everyday egoists. It seems to me that ordinary language simply does not square nicely with the view that the only things that move us are our own beliefs and desires.

The second argument for the view that everyday altruism requires us to relax proposition (i) rather than (ii) rests on conceptual analyses and some scattered empirical observations. Recent research in developmental psychology has shown how very basic the understanding of the purposiveness of other people's behavior is in human cognition. Toddlers are experts in identifying other people's goals, and they are exceptionally successful at this task long before they develop any theory of mind (Tomasello 1998). Infants can grasp what another person's behavior is supposed to achieve long before they pass false belief tests, i.e. long before they have an idea of the other agent as acting on beliefs which may differ from the infant's own.

Thus it seems that the understanding of other people's *intentions* (in terms of what this person tries to achieve) is basic to the conception of another person, not the other way around. Now there seems to be some empirical evidence that links the understanding of another person's intentions to action tendencies of the observer. I will not go into the controversy that revolves around simulation theory and the role of mirror neurons here; instead, let me only point out that a tight conceptual link between understanding on the one hand and acting on the other is at the very origin of the history of the concept of empathy. Theodor Lipps (who can be considered the father of the concept even though others used it before him) observed such phenomena as people in an audience who, sitting in their seats and watching a tightrope walker, seemed to compensate the acrobat's imbalances with movements of their own bodies. Empathy is, Lipps claims, "internal co-action" ("innerliches Mittun";

[10] We should be careful here not to be too quick to jump from such descriptions to ascriptions of pro- (or rather: con-) attitudes. A person might well have the desire to help others, but be helplessly unperceptive or slow on the uptake as far as other people's goals are concerned. Also, it seems obvious that individuals who *do* perceive other people's intentions are not entirely *passive* with regard to the degree to which they let other people's attitude guide their behavior. One might well direct oneself to be more or less accommodating towards other people's intentions – a training which might be guided by one's desire.

Lipps 1903).[11] Understanding isn't originally motivationally neutral. There is an action impulse that flows *directly* from the very understanding of the other agent's behavior, and it is aimed towards the same goal. Thus the very act of understanding provides some motivational steam in and of itself, and it seems that much of the behavior described above operates under this steam. It seems that in the act of understanding, the cognitive (or theoretical) and the conative (or practical) components are internally intertwined.[12] This is shrouded by the fact that it is possible to disentangle the two elements. It is possible to *suppress* the sympathetic components involved in empathy, and to understand other people's behavior without any cooperative impulse, or even to combine empathy with *antipathy*, such as in the notorious case of the cruel person who gloats over his victim's suffering. Some authors have taken this case to prove that there is no internal link between empathy and sympathy at all (Scheler [1912] 1954). Entirely un-sympathetic empathy, however, is *derivative*. Take the case of somebody getting stuck in the middle of a sentence. There is an immediate impulse in the listener to finish the speaker's phrase, perhaps even against his or her own wishes. Similarly for the case of an elderly person struggling to lift her suitcase on the luggage tray. To understand what she is struggling to achieve already means to have an *impulse* to lend her a hand. It does not seem necessary to assume any extra desire to be a helpful person, or a desire to be kind to other people, or some such antecedent motivational state on the altruist's part.[13]

From these (admittedly weak) clues emerges a picture of the intentional structure of the behavior in question that is very different from other-directed desires explanations. It seems that the perception of the beneficiary's intentions is much more closely linked to the benefactor's behavior than the other-directed desires explanation has it. If the suggested line of interpretation is correct, it seems that everyday altruism does not require a particular desire from the benefactor's part for the purpose of a motivational explanation of her supportive behavior.[14] If a person moves

[11] For a reconstruction of Lipps' account in the context of the current debate on the topic see Stueber 2006.

[12] The current discussion on "besires" (a word that combines "belief" and "desire" to refer to mental states that seem to have both world-to-mind and mind-to-world direction of fit, but are different from declarations) seems to come close to the phenomenon at issue here. The paradigmatic case in the besires literature, however, is moral judgment, which is very different in structure from empathy.

[13] In current research on altruism, the assumption that there is a tight link between the understanding of other people's intentions and action tendencies receives strong support from Dan Batson's research. Based on his experimental work, Batson has developed the "altruism-empathy hypothesis" that states that empathy and altruistic action go hand in hand (cf., e.g., Batson 1994).

[14] What is important, however, is the difference between the agent's *not having a desire* to x in terms of the agents not having a pro-attitude towards x, on the one hand, and the agent's having what one might call a "con-attitude" towards x, on the other. For this point, and for the most compelling case for other-motivated behavior in the received literature, cf. Paprzycka (2002). If there is a con-attitude, or some other conflicting desire from the agent's part, other-motivated behavior will not ensue. But this does not mean that there needs to be an additional pro-attitude in addition to the perception of the other's intention for the agent to behave as she does, if the con-attitude is lacking.

aside on a park bench, she might do so simply because she *sees* that the other person wants to sit down (cf. Paprzycka 2002). The motivation is in the perception of the other's desire rather than in anything *she* wants. In this sense, the intentional structure of the behavior in question would indeed not be rooted in the agent's own desires, but in the other's. Insofar as this is the case, we may label everyday altruistic behavior *other-motivated.*

§28 Another Solution to the Paradox

I am well aware that I have not presented any conclusive evidence for the assumption that there really *is* such a thing as other-motivated behavior. So far there is no proof that any of the cases I have discussed *cannot* be explained in terms of other-directed desires; all I have are reasons for doubt, coming from folk psychology, the theory of empathy, and, as we shall see in the following, from some strands in psychological research. I think, however, that these clues are strong enough to give the heterodox interpretation of the behavior in question – i.e. the interpretation that takes the behavior in question to be other-motivated rather than other-directed desire motivated – the benefit of the doubt. I will not go further into this issue here, but instead come back to the initial problem. In this third and concluding part, I will address a purely hypothetical question: *if* other-motivated behavior really existed, could it be altruistic *action*? If we reject other-directed desires explanations, how, then, can the paradox of altruistic action be resolved?

Let me start by stating again the basic problem, which is to show how an individual's other-motivated behavior could instantiate his or her own action. This is an important problem to solve, because one reason why most philosophers do not even think of the possibility of other-motivated behavior seems to be the belief that this problem cannot be solved. If a benefactor's behavior were indeed to be other-motivated, it seems that it would be altruistic in the naïve, strong, intuitive sense mentioned at the beginning of this chapter, leading right back into the paradox of altruistic action.

The argument for this view is the following. According to the orthodox view that is most famously expressed by Davidson, "*R* is a primary reason why an agent performed the action A under the description *d* only if *R* consists of a pro attitude of the agent toward actions with a certain property, and a belief of the agent that *A*, under the description *d*, has that property" (Davidson 1963: 687). Again, this does not mean that no other pro-attitudes than the agent's own can shape his behavior. Rather, the claim is that other people's pro-attitudes are taken into account only insofar as this is what the agent *wants*, i.e. insofar as there are other-directed desires in which the motivational explanation of an action bottoms out. This is at odds with other-motivational behavior. Since an adequate intentional interpretation of the behavior in question would not bottom out in any of the *benefactor's* desires, but in the *beneficiary's*, it seems that, according to the standard notion of action, the behavior in question would have to be attributed to the *beneficiary* rather to the benefactor

himself, since it is the beneficiary's pro-attitudes in terms of which sense can be made of the benefactor's behavior, not any of the benefactor's own. This, however, is in conflict with the *deep-seated notion that, under normal circumstances,*[15] *each single individual's behavior constitutes that individual's own action, i.e. that each individual is an agent*. This is a notion that should not be dropped light-heartedly. It seems it cannot be dropped without thereby excluding some members of the class of agents, and that does not seem right. The idea of individual agency is basic for our most elementary practices of mutual score-keeping in ascribing commitments and entitlements, rights and responsibilities. The importance of this notion can be further emphasized by pointing out its connection to core normative notions. Dropping the idea of universal individual agency (or individual intentional autonomy, as I shall call it) may result in taking those individuals who play subordinate roles in social life to be their superior's extended body, rather than agents' of their own – remember that Aristotle's slave is defined by the fact that he or she is his or her master's instrument. In this case, all sorts of authoritarian, patriarchal or even worse ideas about which individuals do and which do not count as agents in their own right seem to be licensed. The question of who counts as an agent becomes a matter of societal power distribution. And this simply does not seem right. Even convinced Foucaultians are reluctant to deny the powerless a claim to their own agency.

This is one of the reasons why we should not let go of the idea that each individual's behavior instantiates his or her own action. But let's focus for a moment on why this basic and deep-seated assumption seems to be at risk here. If other-motivated behavior were to exist, it seems that the expression "lending one's hand" would have to be taken quite literally. It might appear that whoever lends his or her hand would thereby cease to be the agent behind his or her hand's behavior; that behavior would then have to be attributed to the agency of the person to whom it is lent out, as it were. The role left to the benefactor would indeed be no more than that of a mere *instrument*, or *organ*, of the beneficiary's will, not that of an agent in his or her own right. And this seems utterly implausible, even more so than in the case of submission to power. It is simply not true that people lending their hands lose their status as the agent behind their hand's behavior. They are still held responsible, and even from an internal perspective, it is implausible to assume that lent hands become parts of the other's extended body. Lent hands do not move on the other's remote control. So there seems to be no reason not to hold on to the idea that, insofar as an individual's behavior instantiates any case of action at all, it has to be *the individual's own action*. If we hold on to this assumption, however, the verdict against other-motivated behavior seems to be spoken, for it appears that an individual's behavior can be an action only if it is possible to make sense of it in terms of that individual's *own* intentions and desires. Thus it seems that there can be no other-motivated behavior.

I think that this line of reasoning is flawed. There is a way to reconcile the benefactor's own agency with the possibility of his or her behavior's being

[15] "Normal circumstances" exclude such cases as reflex behavior, coughing, sighing, blinking, where such behavior is not purposefully caused by the agent as a part of his action.

other-motivated. The apparent verdict against other-motivated behavior is due to a confusion in our standard conception of agency that needs to be sorted out. It is necessary to distinguish the following two claims, which are usually lumped together:

1. Individual Intentional Autonomy: *Under normal circumstances (barring certain cases of reflex behavior and pathological dissociations between will and action*[16]*), an individual's behavior is to be interpreted as his or her* own *action.*
   I take it that this assumption is at the heart of our standard conception of agency, and I suggest that for some of the reasons I mentioned above, we should hold on to it. The problem, however, is that on a regular basis the claim of individual intentional *autonomy* is lumped together with a further and much stronger claim:
2. Individual Motivational Autarky: *Any motivational explanation of an individual's behavior has to bottom out in some of that individual's own desires*.

Whereas intentional autonomy is a thesis on which agency is instantiated by a given complex of behavior, *autarky* is a thesis about the motivational resources on which agents may draw. As the thesis states that the only resources are the agent's *own*, it claims that individuals are something like closed intentional economies. Therefore I label this thesis "motivational autarky".

Let's have a closer look at this second assumption before examining how it relates to the first. What does "bottoming out" mean in this context? The assumption of individual motivational autarky does not mean that other people's pro-attitudes couldn't play any role in the explanation of an individual's behavior. As stated above, nobody denies that people do sometimes act on other people's desires or intentions, but it is claimed that they do so if and only if they have *a desire of some sort to do so*, i.e. *on the basis of* an other-directed desire *of their own* in which an intentional explanation of their behavior has to be based. So individual motivational autarky is compatible with altruism in terms of other-directed desires-explanations. It is not, however, with heterodox explanations. For the whole point of these explanations is that they appeal to the beneficiary's pro-attitude rather than to the benefactor's own.

In orthodox explanations of altruistic actions, it is always true that benefactors do what they want because they want it. Orthodox explanations need not thereby deny that sense can be made of the benefactor's behavior in terms of the beneficiary's pro-attitudes. But, in this view, this is only true because there is yet another underlying pro-attitude on the benefactor's part that sustains that link. One can easily imagine more sophisticated cases, in which the boundary between orthodox and heterodox explanations seems to blur. Imagine somebody explaining his altruistic actions with his other-directed desires, but then giving an account of these other-directed desires in terms of a third party's will: "I want to help you because He ordains me to do so."

[16] These include such cases as the *alien hand syndrome,* in which the patient's hand seems to follow an agenda of its own, which even may include the murder of its owner. Other forms of dissociation between will and behavior include *echopraxy* in which patients compulsively imitate the behavior which they observe.

Even though this explanation includes references to other-directed desires, it is heterodox, because it bottoms out in another person's ("His") will (note that, in this case, that other person is not the immediate beneficiary). According to the orthodox view, we have to assume yet another, more basic desire to make sense of this case, e.g. the benefactor's tacit desire to want to do what He wants him to do, or some such additional desire.

So the difference between orthodox and heterodox explanation of the benefactor's behavior is not really a question of whether there are other-directed desires around or not. Rather, the question is where the chain of pro-attitudes quoted in the motivational explanation of the altruistic behavior in question ends (this is what "bottoming out" means in this context): in the benefactor's own pro-attitudes (orthodox explanation), or rather in some other individual's (heterodox). Of course, the simplest case is the most important, where the difference between the two views becomes particularly obvious.

Heterodox explanations are incompatible with individual motivational autarky. Does that mean that they have to fly in the face of the assumption of individual intentional autonomy, too? Let's now have a closer look at the relation between the two assumptions. My thesis is the following: while 2 implies 1, the converse is not the case. Thus there can be individual intentional autonomy without motivational autarky. While heterodox explanations are incompatible with individual motivational autarky, they are in tune with individual intentional autonomy. In other words, the benefactor's behavior need not be taken to be based in any of his or her own pro-attitudes to be interpreted as instantiating his or her own action.

How is this possible? How can a behavior be interpreted as an individual's action without the interpretation bottoming out in that individual's own desires? The answer is this. If one acts on another individual's pro-attitudes, one can form an *intention* to do whatever is necessary so that the other's goal is achieved without having a particular *desire* to do so. In this case, it is true that the benefactor does what he or she *intends*, but it isn't true that he or she does what he or she *wants.* So the *intention* in terms of which the benefactor's behavior is to be made sense of is the benefactor's own, but not the *desire* in which it is motivationally based. Thus there is still a sense in which we *want* to do what we intend to do; intention is a motivation-encompassing attitude (Mele 2003), and it remains so. The constitutive desire is ours. But not so for the motivating desire. On occasion, it is not the case that our constitutive "wanting to A" is motivated by any of our own desires. Our wanting to A may well be motivationally explained by other people's desires. This does not displace our own agency. Thus the benefactor is not intentionally autarkical, but he is intentionally autonomous. It's not that the beneficiary acts *directly* through the benefactor's behavior, as if on the other's remote control.[17] The benefactor still does what he or she intends (and thus constitutively wants) to do *him- or herself*. This case shows how one's behavior can still be interpreted as *one's own action*, even though *the intentional interpretation of one's behavior does not bottom out in one's own individual pro-attitudes*.

[17] For a different heterodox view on the matter cf. Paprzycka (2002).

In other words, and to add a new species to the philosophical literature: I'm not the other's *motivational zombie* if I move aside to make room for him on the park bench without having any corresponding other-directed desire of my own. Motivational zombies are individuals whose behavior does *not* constitute *their own* actions, but rather another individual's, or a group's.[18]

Intentional zombies abound in Sci-Fi, in early accounts of hypnosis and mass-suggestion, in self-reports of schizophrenics, and in some of Al Mele's recent books. But do intentional zombies exist in the real world? It is clear that people can be manipulated into doing all sorts of things; but it is important that only the strongest type of manipulation would amount to intentional zombieism.[19] It is sometimes said that under hypnosis something closely approaching intentional zombieism can be effectuated.[20] I have no knowledge of such cases, and it seems that if they occur, they are limited to very short behavioral sequences. It is clear, however, that everyday altruists are not motivational zombies. My moving aside is still *my own action,* but the intentional resources going into it – the desires motivating my behavior – extend beyond my own pro-attitudes. My intentions are linked to the other's pro-attitudes in much the same way in which normally, my intentions are linked to my own motivating desires. Just as I normally form the intention to sit down on the basis of my own desire to rest a little, without needing another, yet more basic desire to do what I want to do, I can form the intention to move aside on the base of the other's desire to sit down, without there being an additional desire to do what the other wants. Nothing about this structure is particularly mysterious. And it does not affect the agent's individual intentional autonomy.

Again, I do not claim that this is what's actually happening; all I claim here is that it is not *necessary* to abandon the principle of individual intentional autonomy to accommodate other-motivated behavior. All that needs to be abandoned is the dogma of individual motivational autarky.

---

[18] It might be noted in passing that this type of philosophical zombie seems to be somewhat closer to the voodoo idea of zombieism than the "phenomenal zombies" that abound in the philosophical literature, as the distinguishing feature that marks out zombies from other creatures does not primarily seem to be that zombies do not have a consciousness, but that they do not have a *will.*

[19] If the students in a class agree to keep quiet or show a friendly face when the teacher is on the front left side of the room, and chatter or look angry when she is on the right, and if they don't do this too conspicuously, they will soon find their teacher sticking to the left all the time, probably without having any idea about the scheme. But this does not bypass the teacher's agency. It isn't the case that the teacher didn't *intend* to do what she did: upon questioning, she will probably answer that she "likes it better" to be standing on the left side. The problem is, that she does not know that she wanted to do what she did because other people manipulated her into wanting it. So the teacher is very far from being the class' intentional zombie.

[20] It is claimed that in deep hypnosis people may be instructed to show some nonsensical behavior, and that after having woken up, they do what they have been told without having a memory of the instruction, and without having the slightest clue of why their hand suddenly moves. From the internal perspective, the phenomenon is that of the *alien hand syndrome* mentioned above, only that this time there is another person's will behind the behavior, which amounts to intentional zombieism.

If this is true, if individual motivational autarky is no essential conceptual ingredient of action, the question arises: how come it is always lumped together with the idea of intentional autonomy? *Why do we tend to mix up the idea of being the agents* responsible *for our behavior with the very different idea that, in the last resort, only* our own desires *are fit candidates to make sense of our behavior?*[21] In short, my answer is this: it's because in *our culture*, at least, motivational autarky describes the way people are *supposed to be* (and see themselves). Being the one and only ultimate motivational source of the intentional infrastructure of one's own behavior is not a conceptual feature of agency, but it is a very strong normative ideal.

This claim needs some explanation, especially in view of the thoroughly *positive picture* of other-motivated behavior that I have given so far. Think of the person holding the door for another person, or the person moving aside on the park bench, or the one assisting an elderly person with her luggage. How could all these spontaneous niceties, these acts of kindness ever be *in conflict* with any plausible normative ideal? Given the list of examples at the beginning of the chapter, it might even be tempting to explain other-motivated behavior as a kind of *internalization* of social norms. After all, what all the people in the above-listed examples are doing is just being polite. It is important to see, however, that while in most cases other-motivated behavior can be seen as "pro social", there are other cases in which it goes *against* the norms of proper conduct. Thus we might be required to *suppress* the impulse to finish a sentence for a person who is struggling with stuttering – out of simple respect for that person's integrity. And, in education, it is very often *against the norms of proper conduct* to let oneself be carried away by one's other-motivated impulses, because children need to be given the opportunity to exercise their own agency. In most cases, social norms do favor other-motivated behavior. In some other cases, however, this is not true. In these cases, it is not just important that people's goals are achieved; what's even more important is that people can achieve their goals *themselves.*

While a person's explaining her behavior in terms of another person's intentions is frequent in everyday talk, we tend to press for "deeper" explanations, and even to react *embarrassed*, if a person fails to come up with one of her own desires in explanation of her behavior. *People, we seem to think, shouldn't be doing things just because other people wanted them to be done.* People should be *self-reliant* about their own goals, and not be a motivational pawn in other people's play. Thus motivational autarky seems to be part and parcel of our idea of full-blown selfhood and personal identity.

A vivid illustration for the value of motivational autarky and the dangers of other-motivational behavior is provided by Stanley Milgram's famous experiments. As most readers will remember, Milgram's test subjects – perfectly decent ordinary people from a suburban milieu – proved to be willing to administer potentially deadly electroshocks to innocent others, just because they were told to do so by

[21] This conjunction of the idea of individual intentional autonomy and individual intentional autarky is particularly strong in Philip Pettit's *Common Mind* (1996), where he defends intentional psychology against collectivism under the label "individual autarchy".

some authority figure within an experimental setting. There were neither financial incentives nor sadistic inclinations involved. So how come those people did what they did? Milgram himself explains his results by what he calls an "agentic state" (Milgram 1974). An agentic state, Milgram says, is a condition in which a person sees herself as acting on another person's desires rather than his or her own. In Milgram's view, his test people's perception that the motivational base of their behavior is alien to their own psyche explains why their conscience is outflanked by their behavior: only behavior that is motivationally based on the agent's own desires is subject to moral control. While the behavior of Milgram's test people is very different from other-motivated behavior as characterized above in many respects, his concept of the agentic state captures very nicely the central feature of heterodox explanations, according to which people sometimes do what they do, not because of anything they want themselves, but because of what other people want.

As far as I can see, Milgram does not give a clear answer to the question of whether the agentic state is an actual fact about the motivational structure of agency, or whether it is just the delusional self-image of people acting under the influence of authority. However, he seems to be somewhat biased towards the latter reading when he reproaches his compliant subjects for being unable to keep their own act together and assuming responsibility for what they did by claiming that they acted on none of their own desires (this is particularly obvious in Milgram's discussion of Elinor Rosenblum, which is one of the case studies in Milgram's book). In these passages of Milgram's analysis, the agentic state seems to be no more than the test people's attempt to protect their self-image from what they did by blaming the events on the authority. Also, Milgram depicts the agentic state as an *unusual* condition, one that requires the presence and massive influence of authority. And, as is well understandable from the setting of his experiments, he portrays agentic states as morally utterly condemnable. Thus the normative ideal of motivational autarky becomes very clear in Milgram's depiction of the fatal consequences of agentic states.

By contrast to Milgram, and in light of the above examination of the structure of motivationally non-autarkical behavior, I propose to consider three things: first, it seems worthwhile not to dismiss the possibility that experiences of agentic states might be more than just cover-ups used by agents to keep their self-image clean of their wrongdoings. The alternative is to see the self-perception involved in agentic states as referring to actual matters of fact about the motivational structure of behavior. Second, we should consider the possibility that agentic states might be a rather normal condition that permeates much of our everyday live and need not be limited to the presence of authority figures, and which, third, may lead to morally disastrous consequences under conditions such as those examined by Milgram, but can also be very beneficial under such circumstances as to be found in airports and railway stations, among many other places.

Considering the wide range of behavior at stake here, it seems difficult to answer the hypothetical question of whether or not we should uphold the idea that people shouldn't do things only because other people wanted them to be done if it turns out that motivational autarky is in fact a value and not a conceptual feature of action.

I will not pass any judgment here. What is certain, however, is that we cannot even discuss the question of whether or not motivational autarky is indeed an ideal worthy of defense, if we continue mixing it up with intentional autonomy. Because intentional autonomy is a constituent of *any* action, it is not to be changed. By contrast, motivational autarky might turn out to be a cultural ideal, which we may or may not want to uphold.[22] In either case, it is important to distinguish the two.

I conclude with a brief summary of my line of argument and with a remark on the bearings of my results for the economic model of human behavior. The paradox of altruistic action consists of two propositions, which seem to be intuitively plausible, but mutually exclusive. The first proposition is that, for a complex of behavior to be an action, it has to be based on the agent's pro-attitudes. The second is that, for a complex of behavior to be genuinely altruistic, it has to be made sense of in terms of other people's interests rather than in terms of any of the altruist's own. The received literature tends to solve this paradox by relaxing the second proposition and by allowing altruistic action to be based on a particular type of desires. I argued that, while this solution is convincing for a wide range of cases, there is a particular class of altruistic action with regard to which it does not seem to work well. I defined everyday altruism as spontaneous cooperative behavior in low-cost situations, and I provided some clues that seem to indicate that, in order to accommodate such behavior, we might be forced to relax the first proposition. In the last section, I distinguished a weaker from a stronger reading of the first proposition, and I labeled them individual intentional autonomy and individual motivational autarky. There is no conceptual necessity to assume that agents need be motivationally autarkical.

I left the decisive empirical question of the role of motivational autarky in human interaction open, and I should say a word on how the question could be decided. Throughout the chapter I took desires or pro-attitudes to be whatever rationalizes an agent's behavior, given his beliefs. I take rationalization to be a matter of motivation rather than justification. Even though there are unconscious beliefs, I take a special epistemic authority to lie with the agents' themselves, so that the question of whether or not an agent has a desire is answered by the agent's assent under suitable conditions. Unconscious desires are such that they become conscious under suitable conditions (cf. Searle 1983). In light of this view, the question concerning folk psychology becomes particularly important, at least if we assume, as seems reasonable to do as long as there is no evidence to the contrary, that ordinary language is not permeated by some "false consciousness"; hence the special emphasis on folk psychology in the above.

To wrap up the argument, a short remark on the consequences for the economic model of behavior. In the critique of economic thinking, the idea that the link

[22] My claim that intentional autarky is a cultural ideal does not entail that it is necessarily culturally relative, as there might be – and actually are, I would like to think – values that are upheld in all cultures. Considering the ambivalent role that individual intentional autarky plays in our lives, however, I would expect that it is stronger in some cultures than in others. This, however, is an empirical question.

between people's desires and their choices might not be as tight as is assumed by the orthodox account has played a mayor role. The analysis of phenomena such as weakness of the will have helped to cast serious doubt on the assumption that people's behavior adequately reflects their desires. The argument delineated in this chapter hints at yet another way in which the link between what people want and what they do might be more complex than is normally assumed. It might be that, even where behavior does reflect an individual's motivating desires, the desires in question might not be the *agent's own*, after all. Where agents are not motivationally autarkical, i.e. where people's psychologies are permeable to other people's motivations, and where there are relations of spontaneous cooperation or mutual identification between agents, including some forms of influence, power and authority, it might not be so easy to say whose desires an individual's choice reveals. The argument developed in this paper strongly suggests that people can and should be seen as *agents* even if they are not motivationally autarkical.

In his powerful and trenchant critique of rational choice theory, Amartya Sen has claimed that committed agents may act on other people's goals without making them their own (cf. Sen 1985; Peter and Schmid [eds] 2007). While most of Sen's critical points are widely accepted, this particular and uniquely radical claim has been met with considerable skepticism (cf., e.g., Pettit 2005); it has been argued that any violation of the assumption of self-goal choice would simply *displace* the individual's agency. In an earlier paper, I have argued that Sen's claim does make sense as far as *shared desires* are concerned (Schmid 2005a), at least as far as shared desires are irreducible to interrelated individual desires. In light of the above considerations, I would now tend to go even further and argue that there is a sense in which a desire does not have to be an individual's own, or jointly held with other individuals, in order to motivate that individual's action. Still, an individual needs to have his or her own goals in order to be an agent. But, as far as motivational autarky is not part of the concept of agency, the motivational base of the intention that defines the goal need not be any of the individual's own desires for her to be an agent.

It is tempting to think of the relation between other-motivated action and other-directed desire-motivated action as a matter of some switch of frame of mind, in which, in a given situation, one may either act spontaneously and unthinkingly on other people's desires, or decide to take the time to think about the matter and follow the orthodox route by basing one's decision on one's (egoistic or altruistic) pro-attitudes. I admit that this may very often be the case; but Sen's discussion of the structure of committed action seems to point towards the possibility that other-motivation might not only be a matter of unthinking low-cost cooperative reflex behavior, but extend to fully conscious and deliberate choices in which the stakes are great.

Part III
Engaging the ‘Classics’: Four Critical Readings

Chapter 9
Martin Heidegger and the 'Cartesian Brainwash'

Towards a Non-individualistic Account of '*Dasein*'

Intentionality is usually taken to be a kind of solitary object-representation in the mind of individuals. That might explain why intentionalist approaches are so often criticized for being anti-social. To choose intentionality as a starting point of philosophical analysis necessarily seems to lead to a rather under-socialized picture of our cognition and agency. It is a widely held opinion in current philosophy that it takes a radical shift of paradigm to correct this picture, a shift from intentionality to communication (cf. e.g. Habermas 1987), from representation to discursive practices (Brandom 1994), from the analysis how mental phenomena refer to the world to the analysis of the normative social practices and institutions that make utterances *count as* expressions of intentional states such as beliefs or plans for action. Some German philosophers – among these Jürgen Habermas and Karl-Otto Apel – call this shift of paradigm the *intersubjectivist turn*.

This correction of the under-socialized intentionalist picture of the mind, however, comes at a price – or so I shall argue. By reducing the ontological question of *what there is* to the question of the normative practices and institutions within which something *counts as* something, intersubjectivism loses sight of the *objective* aspects of intentionality. If it is true that our mind is not to be understood without taking notice of the social customs, norms, and institutions within which we think and act, it seems no less important to be aware of the fact that we measure our cognitive or practical intentional states not only by social propriety, but also by objective truth or instrumental success. And there is no 'prestabilized harmony' between the two: there is no guarantee that in a given instance the communal practices and institutions within which we think and act help us to see the world as it is. Whereas the former is a question of social normativity, the latter is not. Simply put, social normativity cannot account for all of our cognitive and conative competence. So it seems that we are caught in a dilemma between an under-socialized (intentionalist) and an over-socialized (intersubjectivist) concept of mind. Against this background, I find those recent attempts particularly appealing which try to accommodate sociality in a revised and widened theory of intentionality instead of discarding intentionality as a starting point of philosophical analysis. My conjecture is the following. If most received accounts of intentionality take intentionality to be a kind of solitary object-representation in the mind of individuals, this is the effect of what Annette Baier calls the 'Cartesian Brainwash', and not of some conceptual limitation of intentionality as such. The problem is not intentionality, but rather our standard view thereof.

H.B. Schmid, *Plural Action: Essays in Philosophy and Social Science*, Contributions to Phenomenology 58,

Upon closer inspection, however, it becomes apparent that the effect of the Cartesian brainwash even extends to our current theories of collective intentionality, we-intentions, or shared cooperative activity. Most of these theories stick to the assumptions of methodological individualism for fear that intentionality which could not be reduced to intentionality of single individuals in one way or another would then have to be attributed to some kind of a single group mind over and above the minds of the participating individuals. It's for fear of the group mind that most theorists of collective intentionality endorse one or another version of individualism. I have argued above that there is no reason to be afraid of the group mind (cf. Chapter 2). The specter of the group mind arises from the mistaken Cartesian assumption that *cogitationes* require one single cognizing mind, one single ego – which leaves the collective mind as the only alternative to the individual ego. Thus the anti-collectivist reservations of current theories of collective intentionality and the view of intentionality as a monological matter seem to have the same source. It is the Cartesian brainwash that prevents us from seeing that it is not only single minds, but also interrelated individuals (in terms of "minds-in-relations") who have intentions.

In the following, I shall discuss these issues within an interpretation and critique of Heidegger's concept of *Being-with* (Mitsein). The reason for this apparent detour is that Heidegger's analysis of *Dasein* and its critique contains all the relevant issues and controversies in a nutshell, as we shall see. In many respects, Heidegger's views on the matter are rather ambivalent. Without doubt, the traditional view has its point in claiming that Heidegger's analysis of *Dasein* is just as 'monological' as any of the received theories of intentionality, and that Heidegger, too, cannot account for the social preconditions of cognition and action, thus stepping into the intentionalist trap in which Husserl's transcendental phenomenology ended up before him (cf. e.g. Theunissen 1964). Some American philosophers, however, have recently pointed out that there are some traits in Heidegger's thoughts which cannot be fitted easily into the traditional view. These interpreters have started to portray Heidegger in quite a different hue, depicting his analysis of *Dasein* as a sort of proto-intersubjectivist thought. The ambiguities stretch even further. In most respects, Heidegger's concept of *Dasein* has deeply entrenched individualistic features; but then again, Heidegger at times also seemed to subscribe to a collectivist point of view, calling not the individual, but the total of 'the people' a *Dasein*.

Yet there is more to Heidegger's view on the sociality of *Dasein* than these ambiguities and ambivalences. Above all, Heidegger's analysis of *Dasein* – not so much in its exposition in *Being and Time* as in some of his lectures around that time – includes some elements of a theory of collective intentionality (or inter-intentionality, as I shall call it) that goes beyond subjectivism and intersubjectivism and beyond the alternatives of individualism and collectivism. It is this trait of Heidegger's thoughts on the sociality of *Dasein* on which I shall try to shed some light in the following. I will first turn to the most basic ambivalence in Heidegger's analysis of *Dasein* and the dilemma of the received interpretations (§§29–30 below), before gathering some elements of a solution and relating my reading to current collective intentionality analysis (§§31–32).

§29 The Rift in Heidegger's Concept of Everydayness

In division I of *Being and Time*,[1] Heidegger introduces the term *falling* (Verfallen). *Falling* characterizes *Dasein*'s everydayness – i.e. *Dasein* as it is "at first and for the most part". On the one hand, Heidegger describes *falling* as one of *Dasein's* positive or structural features. As *fallen*, *Dasein* is *concrete*; it is involved with the world and with other *Dasein*. In this sense, the concept of *falling* expresses Heidegger's fundamental insight against Husserl: there is no pure ego and no pure reflection that is logically prior to the 'naïve' straightforward-attitude of everyday life, there is no "subject" over and above the practical involvement with "the world". This is the line in Heidegger's analysis of everydayness that was taken up by Gilbert Ryle (cf. Schmid 2003c: 156–9) and developed into *ordinary language philosophy*. Here, the ordinary – in Heidegger's term: the *falling* – has a thoroughly positive meaning.

On the other hand, Heidegger is not only Husserl's critic, but also his student. As such, he does not simply discard Husserl's reservations against the everyday "natural attitude". This is shown by the fact that, in spite of Heidegger's repeated claim to the contrary, there is always a slight note of depreciation in Heidegger's remarks on the *falling*. In *Being and Time*, the *falling* plays not just the positive role of an integral part of *Dasein*'s existence. It also plays the role of a fatal tendency of *Dasein* somehow to "misunderstand" itself and to live past its own life, as it were (Let's call this the *negative* or *inauthentic* role of *falling)*. Even though these two roles are not strictly incompatible, they make, as we shall see, Heidegger's analysis of everyday *Dasein* at least ambivalent.

To introduce *Dasein's* everydayness, Heidegger uses two famous pictures. The first picture is the one of the craftsman in his workshop. It illustrates the fundamentally pragmatic character of the world and of our intentionality – a term which Heidegger does not make use of because of its Cartesian and intellectualist connotations. He replaces intentionality with the term *taking care with circumspection* ("umsichtiges Besorgen"). With this reformulation of intentionality as purposive, goal-oriented, instrumental action, Heidegger emphasizes that *Dasein*'s self-reference on the one hand and its "being-in-the-world" on the other are closely intertwined, and cannot be separated.

The other picture shows *Dasein*'s everydayness in a much less favorable light indeed. It is the picture of the *One* (das Man). The *One* – or, as it is sometimes translated, the *They* or *Anyone* – epitomizes the sphere of social normativity in terms of norm-oriented action. As *Dasein*'s norm-orientedness, the *falling* has, following Heidegger, fatal consequences for *Dasein*. It leaves *Dasein* no chance to *be itself*. Whether *Dasein* conventionally sticks to the norms, or purposively breaks them, it always does what *one* does. Instead of being *him- or herself*, Dasein is a mere *One*-self (Man-selbst (Heidegger [1927] 1996: 129)). It is not really *me* who does what *one* does, but merely an exchangeable *anyone*. Social normativity thus seems to

[1] If not otherwise indicated, longer quotes from *Being and Time* are based on the translation by Joan Stambaugh (Heidegger [1927] 1996); the pagination indicated follows the original German edition.

distract *Dasein* from its *own* being, i.e. its *own* possibilities (Heidegger [1927] 1996: 42), making it fall prey to what Heidegger calls *inauthenticity*.

Comparing the two pictures, Heidegger chooses to illustrate *Dasein*'s everydayness – one standing for intentionality, the other standing for social normativity – to the two roles of the concept of *falling* (the "structural" and the "negative" one), the following interpretation imposes itself: obviously, the distinction between the craftsman's shop (the sphere of *taking care with circumspection*) and the public sphere (the *One*) directly reflects the ambiguity of the concept of falling. While the workshop illustrates the positive or structural meaning of the *falling*, the public sphere stands for the negative or inauthentic meaning of the term. Thus one might even think that there is a kind of a *division of labor* between Heidegger's reformulation of intentionality on the one hand, and his account of social normativity on the other. Intentionality qua goal-oriented, instrumental action is assigned the role of the positive, structural sense of the term *falling*, whereas social normativity (qua norm-oriented action) is left with the role of the negative or inauthentic sense.

Whatever one might think of this arrangement, it has one consequence that appears to be particularly dissatisfying. The result is an overt depreciation of *Dasein's* sociality. With a grain of salt one could say that Heidegger's recommendation for everyday *Dasein* is to withdraw from the public sphere of communication and social norms, and to take refuge in his or her lonely black forest workshop, where social relations are strictly functional, i.e. confined to occasional transactions with customers and suppliers (Heidegger [1927] 1996: 105). A consequence of the division of labor between the two aspects of everydayness is that the concept of *Dasein*'s authenticity seems to exclude sociality. This lack of sociality in Heidegger's idea of authenticity has been criticized ever since the earliest interpretations of *Being and Time*. This is not to say that there are no traces of authentic sociality at all. There is, of course, Heidegger's theory of *caring-for* (or *concern*, as the German term Fürsorge is sometimes translated), within which he distinguishes an inauthentic, dominant mode (einspringend-beherrschende Fürsorge) from an authentic, freeing version of *concern* (vorspringend-befreiende Fürsorge; Heidegger [1927] 1996: 122). But, insofar as *Dasein's being-with* (Mitsein) is conceived of in terms of *concern*, it is limited to interaction, i.e. to direct face-to-face-encounters. The theory of *concern* does not answer the question of the relation of authentic *Dasein* to social normativity. Leaving aside for the moment the infamous page 384 of *Being and Time*, authentic *Dasein* does seem to be capable of instrumental action and to relations with concrete others in direct encounters which are mediated by instrumental actions, but it remains utterly estranged from any kind of social norms, customs, institutions, and normative communal practices. Perhaps this apparent lack of a full-fledged concept of authentic sociality is the most often mentioned of the conceptual problems of *Being and Time*.

Thus far the classical view of the problem. This interpretation places the author of *Being and Time* among those philosophers who, because of their preoccupation with intentionality neglect the role and scope of sociality for human cognition and action. By contrast to this, some American interpreters, among them Hubert L. Dreyfus, Mark Okrent, John Haugeland, and Robert B. Brandom, have pursued a different

line of interpretation. Their core claim is the following. In their view, Heidegger's concept of practical intentionality does not in any way disregard or even exclude, but instead *presupposes* the sphere of social norms, institutions, and communal practices. There is, following this reading, no intentionality without social normativity, and there is no instrumental human action that is not, at the same time, action guided by social norms. And Heidegger's theory of *Dasein'*s everydayness is seen as the most important witness for this. The argument for this interpretation runs as follows. In his analysis of circumspective 'taking care', Heidegger shows that the traditional question of epistemology was wrongly put. Just to ask how our subjective consciousness comes into contact with the objective ontological structure of the world means to ignore the fact that in our most basic practical dealings the subjective and the objective aspects cannot be separated, but genuinely – 'always already', as it were – belong together. On this Heideggerian line, all representational theories of the relation between mind and world are accused of ignoring the fact that prior to any mental representation we are *in immediate contact* with the world on the fundamental level of our intentionality. In our practical everyday dealings, intentionality is nothing purely mental. On the basic ontological level, the world is not a "whole of things", which are represented in the minds of rational animals like us, and to which we then ascribe functions within our subjective plans for actions. 'Something' is – epistemologically as well as ontologically – always already 'something *as* something.' Entities are always given to us as situated in the pragmatic connections of our courses of action. Following the American interpreters, this always involves social norms. As they see it, this basic 'taking something *as* something' cannot be (and in Heidegger *is not*) conceived of as a monological activity of single individuals, as the traditional reading of Heidegger's concept of *taking care with circumspection* would have it. Rather, the original bridge between world and mind is here seen as consisting in "public performances which accord to social practices", as Brandom puts it (1992: 48–9). Social norms and institutions rather than monological instrumental projects of action constitute what Heidegger calls the *functionality contexture of the surrounding world* (umweltlicher Bewandtniszusammenhang). With reference to these normative social conditions of possibility of intentionality (in Heidegger's sense), Dreyfus speaks of "social background practices" (1991: 149). Haugeland, in turn, calls this the "common institutional framework" of the "customs and practices of a community" (1992: 38, 32). But whatever it is called, it is always Heidegger's *One* or *Anyone* these authors have in mind. Regarding their claim to interpretative correctness, Dreyfus et al. rely on some of Heidegger's remarks where he does seem to ascribe to the *One* something like the structural role of a condition of possibility of *any kind* of disclosedness of the world, i.e. not just *Dasein*'s inauthenticity (cf. Heidegger quoted in Carman 1994: 219). On an argumentative level, too, this reading of the relation between intentionality and social norms has its strengths. For, if it is along the guidelines of social norms that we learn to interpret our surrounding world, and to use the tools in the way we do, it might seem quite plausible to credit these norms with a constitutive role for the structure of our surrounding world, and for the very functioning of our tools. It is only a very short (if fatal) step from saying "what counts as proper and successful use [. . .] is a function of what

the community itself endorses as such" (Carman 1994: 211) to saying "the very functioning of equipment is dependent upon social norms" (Dreyfus 1991: 154), or "something actually plays a role if, according to the customs and practices of a community, it is taken to play that role" (Haugeland 1992: 32).

We shall come back to this shortly. Let me first point out the effect of this reading on the problem of the rift in Heidegger's concept of everyday *Dasein*. By embedding intentionality in social normativity, the original problem of Heidegger's analysis, the tension between intentionality and social normativity, between instrumental action and norm-oriented action simply *disappears*. For it now seems that Heidegger's concept of intentionality itself is thoroughly imbued with sociality right down to its very base. And since the most basic feature of *Dasein*, its "being-in-the-world", can already only be understood as a kind of a social being-with which is embedded in normative communal practices, it would seem rather pointless to complain about any alleged social deficit in Heidegger's analysis. As this line of interpretation takes the norms and conventions of the *One* to play so prominent a role in Heidegger's reformulation of intentionality, I shall refer to it as the *conventionalist interpretation* in the following.

§30 Conventionalism and Its Limits

In a first step, I shall try to cast some doubt on the conventionalist understanding of the connection between Heideggerian practical intentionality and social normativity. There are, in my view, good reasons to insist on a fundamental difference between the sphere of *circumspective taking care* (i.e. practical intentionality) on the one hand and the sphere of the *One* (i.e. the sphere of social normativity) on the other. The 'classical' view is right in pointing out that there is a deep rift in Heidegger's analysis of everydayness, a rift that is simply overlooked in the conventionalist interpretation. Contrary to both the classical view and the conventionalist interpretation, however, I think there is some good argumentative reason why there *should* be such a rift. Here is why. If we take *circumspective taking care* to mean instrumental, goal-oriented action (see, e.g., Okrent 1988: 41ff.), and if we take the *One* to refer to norm-oriented action, one important difference between these two types of action immediately hits the eye. Goal-oriented action is aimed at (and measured by) *instrumental success*. By contrast, the aim and measure for norm-oriented action is *social propriety*. And these are two different sets of criteria. To put it simply: whether an instrumental action is successful or not depends on the real world whereas in norm-oriented actions, it is *up to us*, as it were, since social propriety is a question of conventions and their interpretation, i.e. of social acceptance. The conventionalist interpretation, claiming that the very functioning of tools depends on social norms, eliminates this distinction and identifies what is an "instrumentally successful use" with what is a "socially proper use" of a *thing at hand* (Zuhandenes).

How could this difference slip anyone's notice? I think there is a simple reason for this. The reason is that Heidegger did not distinguish clear enough between two

types of *things at hand*. There is indeed one particular type of *things at hand* for which the conventionalist interpretation is correct. Examples for this type are traffic signs, banknotes, and chess figures. It is, however, not true for *all things at hand*, and especially not for the paradigmatic kind of *things at hand* Heidegger uses to illustrate his concept of *circumspect taking care*, i.e. for tools like hammers, bridges, or drugs.

The conventionalist theory is true for *things at hand* of the first type. In the case of things such as traffic signs, banknotes, and chess figures, the function is indeed constituted by social norms and conventions. This can easily be made clear by a test John R. Searle introduced to distinguish between what he calls agentive functions and non-agentive functions (Searle 1995: 20–23). The test hinges on the fact that there is a way such *things at hand* simply cannot malfunction. Consider the following question: "Are the pieces of wood we move around on the board when we play chess *really* chess figures, or could it be that we were wrong treating them as chess figures according to the rules of chess?" This is a question anyone who is at least loosely familiar with chess games will immediately recognize as utterly nonsensical. For the function of these pieces of wood as chess figures is *constituted* by the communal practice of playing chess and its rules; there is no real fact of the matter hidden somewhere behind the conventionally ascribed function. Therefore, the conventionalist interpretation is right: the function of these *things at hand* is indeed wholly a matter of social norms and conventions. But this distinguishes such things as chess figures, banknotes, and traffic signs from another type of *things at hand*, i.e. from things such as hammers, drugs, and bridges. This is revealed if we run the test again, this time with an example of this second type. With reference to drugs, it is in no way nonsensical, but indeed a sign of prudence to ask: "Does this pharmaceutical product *really* function as a remedy or does it only *count as* a remedy according to the norms and practices of our medicine?" For it could very well be that the Food and Drug Administration or the norms of folk medicine *ascribe* a function to this substance x it *really* does not have. Without us knowing, it could be a substance that is ineffective or even detrimental to health, even though it passed the FDA or has been in use in folk medicine for centuries. The functioning of *things at hand* of this second type, as opposed to the function of *things at hand* of the first type, is not or not exclusively determined by social norms and practices. Heidegger himself seems to have had no clear understanding of (or simply no interest in) this distinction, even though in the important paragraph on the *handiness* (Zuhandenheit) of signs he discusses not only conventional signs (such as traffic signs), but also signs that are linked to the signified by means of a causal nexus (e.g. the west wind as a sign for a change in the weather) (Heidegger [1927] 1996: 76ff.). The conventionalist interpretation, however, completely covers up the distinction between these two types of *handyness*. Thus, Haugeland explicitly holds that even a substance that is really ineffective can be *at hand*, just as long as it is *believed* to be an effective remedy within a community (Haugeland 1992: 32). It seems to me that the conventionalist interpretation gives up the important option to interpret Heidegger's concept of *circumspect taking care* in a way that is compatible with realism. It would probably be wrong to claim Heidegger for any form of epistemological realism, but it is certain

that there is an anti-idealist side to his thinking about *Dasein* that is simply lost in the conventionalist interpretation. As mentioned above, one of the basic insights of his reformulation of intentionality is that, on the fundamental level of intentionality, the subjective side and the objective side (in Husserl's parlance, the "intentional act" and its transcendent "object") belong together. Thus there is no talking about *taking care with circumspection* without at the same time talking about the ontological structure of the world. *Taking care with circumspection* is not a mode of representation of the world, but rather a matter of *being-in-the-world* in the sense of being in immediate contact with or totally immersed in the world.

Similarly, a Heideggerian view of communal practices should take them to be, not only a matter of what a group *believes*, but *also* a matter of what there *really is*. Therefore, it seems to be inconsistent with Heidegger's view to ascribe to groups and communities ways of *taking care with circumspection*, and uses of *things at hand* that are really ineffective. For this perspective presupposes a distinction Heidegger's concepts are designed to undercut. In this view, *taking care with circumspection* does not appear as a mode of disclosedness of the *world* (as it is), but only as a practice within the framework of a culturally relative world-*view*.

Intentional practices can be instrumentally unsuccessful, i.e. ineffective, even though they conform to the social norms or conventions (imagine the artful treatment of an illness according to some community's practices of medicine that is ineffective as a means to the end). And conversely, such practices can be instrumentally successful without properly following the normative standards of a community (imagine any effective use of a tool for a purpose for which the tool is not designed). In both cases, instrumental success and social propriety come apart in one or the other way. The question is: which of these cases should be seen a as case of *taking care with circumspection?* Let's first hear the conventionalist's reply. Concluding from the above example, Haugeland's reply to the first question seems to be clear: in spite of all ineffectiveness, a practice can be an instance of *taking care with circumspection*. In the latter case, the conventionalist reply seems to be in the negative. Thus Dreyfus states that "a hammer is for hammering and not for opening paint cans" (Dreyfus 1995: 425) (even though, with some hammers at least, this can be done very successfully).[2] Thus it seems that, on this view, whoever deviates from the normative communal practices thereby lacks the proper *circumspective care*, however successful she or he might be.

I think, however, that the reverse replies are closer both to the fact of the matter and to Heidegger's views. As argued above, intentional practices which conform to

[2] When the conventionalist interpreters consider the difference between instrumental success and conformity with norms, they often project it onto the distinction between primates and humans, thereby devaluating "mere" instrumental success as a criterion for the use of tools in higher animals. If a primate uses a stick to fetch some bananas hanging high in the bush, his action is either successful or unsuccessful. Human "taking care with circumspection", on the other hand, seems, following John Haugeland, to be measured by higher standards: it is either "proper" or "improper", something that cannot be said of animal instrumental action (Haugeland 1982: 18). Heidegger, as I see him, would respond: social norms or not, the question about "taking care with circumspection" is whether in the end you get the bananas or not.

social norms but are instrumentally ineffective should not be called instances of *taking care with circumspection*. Intentional practices, however, which are "improper" with regard to the social norms, but instrumentally successful, should not be denied this title. For the latter case, think of a prisoner in his cell who uses his fork and knife to dig a hole in the ground to prepare his escape. The conventionalist interpretation cannot seem to see beyond the fact that he uses these tools improperly. And of course they are right – but only from the perspective of the community whose conventions constitute the normative standards, which define what counts as proper and improper prisoner behavior. It's the warden's view, so to speak. Heidegger, by contrast, allows for a less biased view of the practice in question, and this is done by the very distinction between intentional practices and the sphere of communal norms. As far as the prisoner is not hallucinating, as long as, in other words, 'world' is disclosed in his working towards escape using his spoon as a tool for digging, there seems to be no reason not to credit his practice with the title of *taking care with circumspection*. For circumspective care is primarily a matter of instrumental success, and not of social propriety. Therefore, a distinction has to be made between action of the type of circumspective care on the one side, and norm-oriented action in the sphere of the *One* on the other. The rift in everydayness opens up the room to do justice to such socially improper and idiosyncratic perspectives such as the one of our prisoner.[3]

As far as I can see, the first round of the conventionalist interpretation started with a paper by Haugeland in 1982, followed by contributions to the debate by Brandom (1992) and Mark Okrent (1988) (among others), culminating in Dreyfus' book on division one of *Being and Time* in 1991. With Dreyfus' more recent papers (1999, 2000) and the contributions to the first volume of his *Festschrift* from 2000 (Wrathall and Malpas, eds.), however, it seems that a new round of the debate has started. While the earlier interpretations were focused on division one of *Being and Time*, i.e. the analysis of everyday *Dasein*, the debate has now moved on to division two and the idea of *authenticity* (Eigentlichkeit). Now, the conventionalists see themselves faced with the possibility of authenticity as a form of disclosedness of the world that is not – or not *exclusively* – caught up in social norms. *Dasein*'s authenticity implies a form of critical distance to the sphere of customs, conventions and other rules. While the conventionalists try their best to allow for this possibility, it is not always obvious how the possibility of authenticity squares with the earlier thesis that any form of intelligibility rests on an *a priori* foundation of social normativity. Understandably, the conventionalists were not originally attracted to the idea of

[3] That this is indeed Heidegger's view can be made evident by quoting some of the many passages from *Being and Time* where Heidegger seems to imply that the "they" does not so much disclose as *veil* the world. The conventionalist thus correctly quotes from page 127: "Publicness (i.e. the "they", H.B.S.) initially controls every way in which the world and Da-sein are interpreted", but this is, as he continues, "not because of an eminent and primary relation of being to "things", not because it has an explicitly appropriate transparency of Da-sein at its disposal, but because it does not get to "the heart of the matter" (auf Grund eines Nichteingehens 'auf die Sachen') [...]. Publicness obscures everything, and then claims that what has been thus covered over is what is familiar and accessible to everybody." Heidegger [1927] 1996: 127).

authenticity, and rarely talked about that concept. Where they did talk about the idea of authenticity, they usually twisted the concept beyond recognition. Thus, in his earlier interpretation, John Haugeland claimed authenticity to be no more than the possibility to decide which role to choose in the case of role-conflicts (Haugeland 1982: 23–4) – which was consistent with the idea of the *a priori* of social norms and normative practices, but at the same time meant pulling the teeth of the idea of authenticity, to say the least. Authenticity is more than decisiveness in the handling of occasional role conflicts. Rather, authenticity implies a critical stance on the role-character of one's life as a whole. Or, put differently, authenticity is not just about finding out which role to play, but about living beyond the framework of social roles.

In the second round of interpretation, the conventionalists see themselves forced to acknowledge at least some of this anti-conventionalist edge of the idea of authenticity. Thus, in his new interpretation, Haugeland now seems to allow for some distance from the context of social norms and conventions. Authenticity, he now states, implies something like the capacity to become aware of what he calls the *anomalies* generated by our everyday practices of blindly following the rules and convention of the *One*. Still, however, Haugeland sticks to the claim that the disclosedness of the world is always imbued with social normativity in something like the way in which the scientific experiments designed to test a theory are always based on that very theory itself. Indeed, Haugeland models his reading of authenticity on the theory of scientific progress. What attracts Haugeland to this model is that, in Thomas S. Kuhn's theory, the quasi *a priori* status of scientific theories does not rule out the possibility of "anomalies" within normal science, which become the occasion for scientific revolution. Similarly, the fact that the *One* is a condition of possibility of the world's disclosedness does not rule out the possibility that the conventional practices fail. Transferring Kuhn's theory of science to everyday life, Haugeland now calls 'inauthentic' *Dasein's* tendency to stick to the familiar rules and practices even if there are clear hints for their failure (cf. Haugeland 2000), instead of looking for alternative practices.

Haugeland is not the only conventionalist interpreter to develop a new interpretation of authenticity. Dreyfus comes up with a somewhat different interpretation (Dreyfus 2000). Following up on Theodore Kisiel's reading of *Being and Time*, he now talks about the possibility of an authentic disclosedness of the world that reaches well beyond the sphere of social norms and conventions. Dreyfus' example for the distinction between authenticity and inauthenticity is the difference between an expert and an apprentice in their respective relation to the social norms and practices of their trade. It is in grasping and then blindly following the social norms that we learn using our equipment, Dreyfus observes. By contrast to the apprentice, the expert in her field just *knows how to do it* without necessarily following the rules of a communal practice. Her actions are not norm-oriented in the sense in which the apprentice's are, indeed experts' ways of handling things are typically quite unconventional, for the expert can tell when, how, and why something has to be done conventionally, and when, how, and why it is possible and indeed useful to depart from the conventional standard procedures. Thus the expert finds her own style of skillful coping.

These and similar (Carman 2000) interpretations of authenticity thus open up a perspective on a form of "intentionality" (disclosedness of the world) that is not completely caught up in social normativity. I think Dreyfus et al. are right in moving in this direction. Yet it is an open question whether or not these theories can be made consistent with the earlier reading of the relations between intentionality and social normativity. It is not obvious how the *a priori* of "anonymous social normativity governing intelligibility at large" (Carman 2000: 20) can be made consistent with the idea of a form of "authentic" intentionality, which in one way or another transcends the realm of social normativity. In my view, the conventionalist line of interpretation is caught in a dilemma, being faced with the choice between two equally repellent alternatives. Either the conventionalist line is rejected, or the idea of authenticity loses its anti-conventionalist teeth. The conventionalists follow the second line. It seems to me, that the conventionalist interpretation of authenticity, though moving in the right direction, is thus flawed by the shortcomings of the earlier interpretation of everyday *Dasein*, especially the theory of the relation between social norms and practical intentions in everydayness. All the conventionalists can find in authenticity is what they left out in their earlier descriptions of everyday *Dasein*. I think that the conventionalists should have chosen the first horn of their dilemma instead. But this would have meant to give up the conventionalist stance, and re-open the rift in everyday *Dasein*.

Let's have a closer look at what seems to be at stake here. In the case of *things at hand* such as chess figures and banknotes, an act of intentional *circumspective care* (such as a move in a game of chess, or an act of payment) can indeed be successful if and only if it conforms to the social rules (with the exception of unnoticed cheating, i.e. moves that do not conform to the chess rules, and the use of forged banknotes, which are parasitic cases we do not take into account here). In these cases, there is indeed an *a priori* of social normativity, in the strictest sense of the word, at play. But this does not hold for all *circumspective care*. The relation between success and conformity is radically different in other cases. It is true that our use of hammers, pharmaceutical products, and the construction of bridges usually conforms to the respective social rules and norms governing these practices, too. We usually do all of these things the way *One* does it (or should do it). But still, these cases differ fundamentally from our playing by the rules in moving chess figures. The difference is this: normally (i.e. with the exception of unnoticed cheating) our moves in games of chess succeed (in terms of counting as a move) only if – and only and because – the moves conform to the social norms. This is different with *things at hand* such as drugs. Here, the relation between success and conformity runs precisely the other way around. Successful use is not constituted by social norms, but the norms are constituted by successful use. (The exception to consider here is the case of norms based on wrong assumptions. But this case, just as the case of unnoticed cheating, is parasitic and can be disregarded for the present purpose.) It is only *insofar* and *because* we believe that the norms secure instrumental success that our use of drugs and our constructing bridges conforms to the respective norms. In other words, the difference at stake here is that between constitutive and regulative norms or rules. Things like drugs are not constituted by the social norms regulating their use, but the

social norms have to comply with their successful use, while in the case of things such as chess figures the relation runs the other way around. The conventionalist *a priori* of social normativity holds for the status of things of one kind only. In the case of *things at hand* for which it does not hold – i.e. the status of things such as hammers – we stand in a relation to the success of our actions that is *mediated*, but not *constituted* by social norms. What we have here is a form of intentionality that is not completely caught up in social normativity, but typically remains at a critical distance to conventions. This is not, however, just a feature of some authentic *Dasein* which somehow extends beyond the sphere of everyday life. Rather, it is an integral part of everyday means to have an intuitive and implicit understanding of this basic difference. A person who does not understand that the function of chess figures is constituted by the rules of the game could hardly be called a competent player. Conversely, a person who takes the function of hammers to be constituted by social norms (and therefore concentrates on using the hammer *properly* instead of doing with the hammer whatever is needed to make sure that the nail is driven in), we might perhaps call a continental philosopher, but certainly not a hobby craftsman, let alone a competent hammerer.

Thus the conventionalist interpretations of authenticity in the second round of the debate make the impression of a belated compensation for the earlier misreading of the analysis of everyday *Dasein*, ascribing exclusively to authenticity what seems to be an integral part of everyday *Dasein* as such. But let's come back to our initial problem, i.e. the relation between intentionality and the social character of cognition and action. With respect to this question, the main flaw of the conventionalist interpretation of authenticity is somewhat different. Remember that the conventionalist reading of the first division of *Being and Time* was aimed at eliminating the rift in Heidegger's analysis of everyday *Dasein*. True to the matter and to Heidegger's views or not: the claim was that there is no rift between monological *circumspective care* and the inauthentic sphere of social normativity, because there is no skillful coping outside the *One*. By contrast to this, it seems now that under the title "authenticity" a no less monological form of intentionality is re-introduced by the conventionalist interpreters themselves. For authenticity, understood as a sort of awaking from the dogmatic slumber of conformity (in Haugeland's interpretation), or as outgrowing the straitjacket of social norms in becoming an expert (in Dreyfus' interpretation), does indeed seem to be a rather lonely and monological affair. In both interpretations, authenticity is conceived of as something that explicitly concerns us as *single individuals*. So the initial problem simply reappears. Whereas the conventional disclosedness of the world in the *One* is something genuinely social, the authentic disclosedness of the world is not. Thus Haugeland as well as Dreyfus (at least to some degree) seem to fall back into the old division of labor. The scandal is still there: Heidegger's monologism of authenticity, his inability to allow for any genuinely social dimension of authenticity. And the intuition of the first critics of *Being and Time* still holds: authenticity does concern us not only as single individuals in our solitary *being-in-the-world*, but also in our *being-with* others in our communal lives. The question is: how can we conceive of the social dimension of authenticity?

§31 Joint Action and the Social Dimension of Authenticity

In what follows, I shall try to gather some elements for a different solution to the problem. As I see it, it's not Heidegger's distinction between practical intentionality and the sphere of social norms that is at fault, as the conventionalists think. Heidegger was right in drawing this distinction, and indeed in calling the conventionalism of the *One* inauthentic. To some degree at least, even the conventionalist interpreters now seem to make Heidegger's reasons for this view of the *One* their own. I will propose the following interpretation: the main problem of Heidegger's analysis of everyday *Dasein* is not that he kept the innermost of *Dasein*'s intentionality clear of social normativity, but that he conceived of it in individualistic, indeed atomistic terms. So the solution to the problem lies in a non-individualistic conception of intentionality.

In his introduction of the principal distinction between *Dasein* on the one hand and *beings unlike Dasein* (nichtdaseinsmäßiges Seiendes) on the other, which marks the beginning of his analysis of *Dasein* in *Being and Time*, Heidegger already ties *Dasein* down to an individualistic mode of existence. The basic *existential* is deeply imbued with individualism. "*Da-sein* is a being that does not simply occur among other beings. Rather it is ontically distinguished by the fact that in its being this being is concerned *about* its very being" (Heidegger [1927] 1996: 12). Here, a special kind of *self-reference* is claimed to be the innermost feature of *Dasein*, a view that culminates in the thesis of *Dasein*'s famous *always-being-my-own-being* (Jemeinigkeit) (Heidegger [1927] 1996: 42). It is quite typical of Heidegger, though, that an alternative to his individualistic concept of intentionality and *Dasein* can be found in his own work. Interestingly, Heidegger, in his series of lectures from 1929/30 (published as Vol. 27 of Heidegger's collected works in [1929/30] 1996), Heidegger starts out from the same principal distinction that marks the beginning of the analysis of *Dasein* in *Being and Time*, but he draws it somewhat differently. Here, what distinguishes *Dasein* from other beings is not the way *Dasein* is related to *itself*, but rather the way it is related *to other beings of its kind*. The basic feature of *Dasein*, in other words, is its *being-with*. Here is how Heidegger conceives of the basic difference. It is characteristic of *beings unlike Dasein* that they *occur among* other such beings (i.e. other beings that are unlike *Dasein:* "nicht daseinsmäßiges Seiendes kommt neben anderem nicht daseinsmäßigem Seiendem vor"). *Dasein*, by contrast, does not 'occur among' other *Dasein*. Rather, it *is with* other *Dasein* ("Dasein und Dasein *sind miteinander*") (Heidegger [1929/30] 1996: 85). Based on this fundamental distinction between entities that *occur among* other things and entities that *are with* other entities, Heidegger here introduces *community* (Gemeinschaft) as the most fundamental of the *existentials* of *Dasein* (Heidegger [1929/30] 1996: 145). Most significantly, it is only in passing, as it were, and towards the end of the series of lectures that Heidegger finally turns to the issue that occupies the center stage in *Being and Time*. With the focus on *Dasein's* communal being, the question of how *Dasein* is revealed to itself (Heidegger [1929/30] 1996: 134), which is the dominating topic in exposition of the analysis of *Dasein* in *Being and Time*, loses much of its interest (Heidegger [1927] 1996: 15–40).

This change in perspective is not without consequences. The fundamentally communal character of *Dasein* requires yet another fundamental change in the concept of intentionality. The view Heidegger comes up with is remarkably different from the one put forward in *Being and Time*. Now, he has yet another reason to distance himself from earlier conceptions. It is no longer just the representationalist and intellectualist implications of the traditional concept of intentionality that stand in the way of an adequate understanding of *Dasein*'s *being-in-the-world*. The additional obstacle Heidegger now has to overcome is the *atomistic individualism* of the theory of intentionality. Against this limitation to individuals, Heidegger here outlines a non-individualistic account of intentionality to which I will refer as *inter-intentionality* in the following[4] (as mentioned above, Heidegger himself does not make use of the term intentionality). What he now comes up with is a concept of an engagement with the world that does not depend on social normativity in terms of conventions, normative communal practices or social institutions, but is not an affair of single individual minds nevertheless. It is social, but not conventional. So what is the structure of this sociality? As he does so often, Heidegger starts with a negative characterization. Against possible intersubjectivist or conventionalist misunderstandings,[5] Heidegger here states clearly that the fundamentally communal character of *Dasein* does not mean that it is *only as a member of some community of communication* that *Dasein* is engaged with the world. Rather, the point is that *Dasein* has intentionality not *only* as a solitary individual (even though Heidegger seems to allow for this possibility), but sometimes *shares* intentionality with other *Dasein*. Heidegger does his best to fend off possible individualistic and collectivistic misunderstandings. The following picture emerges. Inter-intentions are neither collective phenomena that are somehow supervenient on individual intentions, nor are they simply social in terms of Max Weber's individualistic concept of social action (social action qua based on individual intentions that are at least partly social in content, i.e. directed towards other individual agents). Inter-intentionality is not a matter of any intentional state that has the other as its object. To put it in Heideggerian terms, inter-intentionality is not a mode of *concern*. Rather, it is *shared intentionality*. In this view, shared intentionality (in terms of acting and experiencing together) does not entail any thematic and explicit relationship to others whatsoever.[6] It is no form of regular individual intentionality, to which some form of knowledge of the other (and the other's experience of the object, and the other's knowledge of one's own experience of the object, and so on) is added, as was first claimed by Gerda Walther in the early phenomenological thinking on social theory (Walther 1923:

[4] It was only after I published this paper that I finally became aware of the true extent to which Heidegger, in the non-individualistic turn described in this section, relied on Max Scheler's insights. Scheler's influence is clouded by the fact that it is not acknowledged by Heidegger.

[5] Heidegger almost seems to address his later intersubjectivist interpreters when he explicitly states that "community" should not be taken as the "only principle" *(alleiniges Prinzip)* of the disclosedness of the world (Heidegger [1929/30] 1996: 146).

[6] Heidegger ([1929/30] 1996: 86–7). This is one of the features of inter-intentionality that Heidegger seems to have taken over from Scheler.

esp. p. 86), and as it is still sometimes claimed in current collective intentionality analysis in general and game theoretic discussions on the topic in particular.[7] The intersubjectivity of inter-intentionality is neither made of social normativity, nor of experiences of or beliefs about the other. So what is it made of, then?

Turning from the negative characterization to the theory, Heidegger introduces an example that is meant to illustrate the phenomenon. The example is the following: two hikers are *carried away* and *dazed* by the sight of the sunset they are jointly watching (Heidegger [1929/30] 1996: 86, 88). They experience the sunset *together*, without their attention being drawn to each other in any sense whatsoever. Significantly, Jean-Paul Sartre uses a similar example to discuss Heidegger's concept of *being-with*. It is well known that in his own theory, which he set off against Heidegger's, Sartre insisted vehemently on grounding sociality in explicit relations between individuals. In this sense, Sartre is at the same time Heidegger's fiercest opponent, and his best interpreter. More clearly than any other of Heidegger's interpreters and critics, he saw that it is possible to read Heidegger in an inter-intentionalist way. Sartre's own example for inter-intentionality is the joint experience of a stage performance. The people in the audience are experiencing the performance *together*, without, however, having any explicit experience of *each other*. The relation between individuals sharing a joint experience, while being essential for the jointness of that experience, is of the non-objectifying kind. *Dasein* is, as Sartre puts it, "non-thetically engaged in a 'we'" (Sartre [1943]1991: 485). In the words of John R. Searle, the intersubjective relation in question is of "pre-intentional" character (Searle 1990: 415). Leaving aside the fact that the Cartesian infallibility does not apply – in contrast to individual intentions, it is possible to be mistaken about our own common intentions[8] – this inter-intentional relation between individuals structurally resembles, indeed is structurally *identical* with Sartre's famous "conscience (de) soi"; just that it is not "(de) soi", but "(de) l'autre", as it were. Like all individualist philosophers, Sartre would like to reserve that innermost of intentionality for the relation of the individual to *itself*, thereby privileging self-reference over all other kinds of relation. But *pre-reflective*, *non-thematical and non-objectifying relations* are not limited to our contact with *ourselves*. Our relations to *each other* are made of the same stuff.

Along this line of thought, a different concept of *Dasein* emerges: *Dasein* as engaged in inter-intentional practices does not have the existential form of an individual *always-being-my-own-being*. Rather, it is a genuinely *communal Dasein*. In more traditional terms, Sartre calls it the *we-subject* ('nous'-sujet) (Sartre [1943] 1991: 498).

Before defending this inter-intentional concept of *Dasein* against Sartre's individualistic criticism, a critical remark on Heidegger is in order. It is remarkable that,

[7] The authoritative text on this topic still is David Lewis' *Convention. A Philosophical Study* (1969); concerning the more recent debate on "common knowledge" see Gilbert (1989: 191ff.); Nozick (2001: 154ff.).

[8] It may well be that I am mistaken about *our* (the gospel choir's) intention to meet for a rehearsal tonight (perhaps I got that wrong), whereas I cannot be mistaken about my individual intention to participate.

in his series of lectures *Introduction to Philosophy*, Heidegger developed a reformulation of intentionality that avoids the shortcomings of individualism. But it is also true that even here he shied away from the consequences of his own insights. Thus even the above-mentioned example he uses is quite telling. The paradigmatic case is a strangely passive group of co-experiencers (the two hikers jointly experiencing the sunset) instead of an active group of co-actors. This might be due to Scheler's influence (who chooses a similar example), but it can also be taken as a clear sign of Heidegger's *profound reservations* about non-individualistic forms of intentionality. Remember that, in *Being and Time*, Heidegger identifies our practical skills as the basic feature of what used to be called intentionality. Intentional stances such as experience and belief are, as Heidegger shows convincingly, *derivative* forms of the intentional. In illustrating inter-intentionality with the example of joint experience, Heidegger seems to keep the innermost of intentionality clear of sociality. Heidegger does not discuss *cooperation* or *joint action*, but joint experience. This is more than just a consequence of some contingent choice of example. Heidegger makes quite explicit in his remarks that inter-intentionality does not disclose *things at hand* (Zuhandenes), but only *things in their objective presence* (Vorhandenes). Thus it seems that the communal form of intentionality takes place not on the basic, but only on a secondary (derivative or even "deficient"; Heidegger [1927] 1996: 61) level of intentionality (cf. Heidegger's convincing analysis of the relation between *things at hand* and *objective presence* in *Being and Time*). Even where Heidegger finally takes the inter-intentional givenness of a *thing at hand* as an example – he chooses a piece of chalk in the classroom – he explicitly *rejects* the idea that the inter-intentional givenness of this thing lies in its *use* within some *joint activity* (Heidegger [1929/30] 1996: 108). Thus, on the basic level of intentionality, everything remains the same. Heidegger is not ready to accept that *taking care with circumspection* is more deeply imbued with sociality than individual instrumental activity, which is instrumentally or strategically linked to other individual's instrumental activity (remember the craftsman's relations to his customers and suppliers). The inter-intentional givenness of things requires that we refrain from using them – that we *let* things *be* (sein lassen), as Heidegger says explicitly.[9]

Also it seems that Heidegger still maintains here what he said earlier in his *Prolegomena to the History of the Concept of Time*. *Things at hand* are either tailor-made for one particular individual's use – or they lie about publicly, as it were, being there for anybody's use, so that whoever uses them thereby turns into a mere *Anybody*, thus entering the inauthentic mode of existence.[10] But this alternative between one particular individual's circumspection and anyone's use is not

[9] Heidegger continues, though, that this "letting be" is not in any way *deficient* as compared to practical use, but lies "before any interestedness" (Heidegger [1928/29] 1996: 102).

[10] Heidegger, in the "Prolegomena to the History of the Concept of Time", speaks of "things at hand" that are "in ihrer eigentümlichen Anwesenheit nicht auf einen einzelnen, auf ein bestimmtes *Dasein* als solches zugeschnitten [...], sondern [die, H.B.S.] jeder in derselben Weise wie der Andere gebraucht ([die, H.B.S.] 'man' im gleichen Sinne verfügbar hat), was für 'einen' schon da ist" (Heidegger 1979: 270).

exhaustive. It is not just within my individual plans of action or within the plan of action of the mere average *anyone* that things are *at hand*. To give an example: if you and I jointly carry a large sofa from the removal truck up to the new apartment on the third floor, the surrounding world of our joint activity – the corners and handrails of the staircase, for example – is not disclosed in the light of my or your *individual* circumspective *taking care* nor in the light of any average individual's *circumspection*. Rather, the surrounding world with all its possibilities, tools and other features is disclosed in the light of *our joint activity*. Inter-intentionality is not just a matter of passive experience. It is, above all, a matter of *joint activity*. Joint action implies a form of disclosedness of the surrounding world, and an individual's participation in a joint action is neither a mode of being towards *his or her own (individual) possibilities* nor a mode of being towards an exchangeable individual *anyone*'s possibilities. In the individualistic sense of the world, common action is neither authentic nor inauthentic. For *Dasein*, as engaged in joint action, neither *chooses* nor *loses* its own individual being (Heidegger [1927] 1996: 42). This becomes particularly obvious if we take into account that the innermost of *Dasein* is conceived of in terms of *possibility* (Möglichkeit). The reason why we have to go beyond the two alternatives of either grasping or covering up *Dasein's* individual possibilities is that, in joint action, we do not act towards our individual possibilities at all. Joint action is about *our shared* possibilities, and these are not merely a sum or an aggregate of the individual possibilities of the participating individuals. There is no way of accounting for shared possibilities in terms of individual possibilities. The reason is not that individuals do not have individual possibilities when acting jointly, but that, in most cases, the individual possibilities they have are *based on* the shared possibility, and not the other way around. To quote a trivial example, it's only within the shared practice of an election that individuals can cast their votes. The possibilities that shape our shared being are the base and frame of many of the possibilities we have as individuals. As observed by Heidegger, possibility is what *Dasein* basically *is*, the very being of *Dasein* is not only *my own being*, but *our common being*. *Dasein* is not – or not *exclusively* – the being of an individual, as the individualistic setting of *Being and Time* makes us believe.

It is true, of course, that joint intentions and actions, too, require individual intentions and actions. There is no such thing as joint action unless there are individuals who act. But this does not mean that shared intentionality and action is something that only emerges (and is thus ontologically dependent on) some underlying individual level. On the contrary, it is the individual intentions and actions involved in collective intentions and actions that are dependent on the collective level of intentionality. Our jointly carrying the sofa up the stairs does not emerge from two independent individual actions. Rather, my individual lifting my side of the sofa above the handrail and your individual slowly stepping around the turn of the stair holding your end of the sofa are intended *as parts* of our joint action, and it's the whole that gives the parts their meaning. Thus my and your individual actions are – in their intentional content, mode, and perhaps subject – to be understood as individual *contributions* to a common intention and action, and are thus dependent on

the shared intention. Within this joint intention, our *contributive actions* are interdependent. They have to mesh in order to be effective as contributive actions. Thus, in the example of two participants, the structure of common action could be viewed as a triangle, its corners being a) our joint intention, b) your individual contributive action and c) my individual contributive action, the line between a and b and the line between a and c standing for the derivative relation between our individual contributive actions and the joint intention – I'm lifting my end of the sofa *because* we intend to move the sofa – and the line between b and c symbolizing the interdependence between your and my contributive action – my "lifting my end of the sofa" constitutes an individual contribution to our common action *only if* it meshes with your individual contribution. These features are currently discussed under labels such as we-intentions, collective intentionality, or shared cooperative activity. The general aim of this debate meshes seamlessly with Heidegger's aim in *Introduction to Philosophy*. The goal is to open a perspective on the genuinely social character of intentionality.[11] Moreover, it seems to me that, to some degree at least, the current debate could learn from Heidegger's thoughts on the matter.

§32 Collective Intentionality: Heideggerian Inspirations

It has repeatedly been noted that the analytic debate on collective intentionality is marked by an individualistic bias (Baier 1997b: 21ff.; Stoutland 1997: 45–74). Part of the reason for this setting seems to be fear of the group mind. Understandably, John Searle finds the assumption of some group mind over and above the individuals an idea that is "at best mysterious and at worst incoherent" (Searle 1990: 406, 1998a: 118). Yet it seems no less incoherent and mysterious to take this as an argument for methodological individualism and even methodological solipsism, as Searle does. The fact that there is no group mind does not mean that all intentionality is to be found in the brains of individuals (as Searle suggests with his reading of methodological individualism), or even that it makes no difference to the structure of the intentionality in question whether those brains are in contact with the real world or just dreaming in a vat (as methodological solipsism has it) (Searle 1990: 406). For it is clear that single brains in vats cannot have collective intentions. An intentional state of the form *we intend* in a single mind, which is not connected to other minds in a suitable way, is not just a collective intention that has somehow gone wrong. It is *no collective intention at all*. This is just another way to say that it is not individuals, but only *individuals-in-relations* that can have collective intentionality. In contrast to Searle, Michael E. Bratman is well aware of this point. "Shared intention is not an attitude in any mind" (Bratman 1999: 122), he says, but an "interrelation" of the "attitudes" of several individuals. Yet Bratman, too, thinks it necessary to endorse individualism (Bratman 1999: 108, 129) in order to avoid the group mind (Bratman 1999: 111). This he does by making the collective intentional activity the

[11] In addition to the titles mentioned below see, for example, Gilbert (1989) or Tuomela (1995).

propositional content of intentional states of the form *I intend*. Thus shared cooperative activity, as he calls it, consists in an *interrelation* of individual intentions of the form *I intend that we J* (J standing for the joint activity that is being planned) (Bratman 1999: 142ff.). This form of individualistic reductionism (Bratman 1999: 109), just as Searle's, has counterintuitive consequences. Annette Baier has argued convincingly that in order to have an intention of the form "I intend that we J" one has to take oneself to be somehow "in control" of what the others do, since one cannot intend to do something one takes oneself to be unable to perform (one cannot intend to spend the afternoon in the library, if at the same time one is aware of the fact that it is closed all day) (Baier 1997b: 15–44). In his reply to this objection, Bratman stated that it is not necessary to take oneself to be in full-blown control of the relevant others. As he points out, it suffices to assume that there is a sufficient chance that one's intention will *influence* others so as to go along with the joint venture. In the paradigmatic case, an expression of one's intention will motivate the relevant others to cooperate (Bratman 1999: 155ff.). This seems to open an alternative to my being in total control of the relevant others in intending that we J, namely our (however latent) collectively intending to J that is just somehow activated by my individual intending that we J. In this latter case, however, it seems obvious that the *collective* intention to J is already presupposed in my *individual* intending that we J.

Looking at this debate, I think there is an important lesson to be learned from Heidegger. It is not to cling to individualism for fear of the group mind. Heidegger once remarked that the idea of a collective mind (to which the individuals belong as mere parts to the whole), though it might superficially seem to contradict individualism, is nothing else than one of the "most dangerous consequences" of the very obsession of modern philosophy with the individual "I" (Heidegger [1936–38] 1989: 321). This, however, did not prevent Heidegger from committing this very same mistake himself. Shying away from the idea of inter-intentionality (that could have filled in the gap, as we shall see), Heidegger had to leave empty the place of authentic sociality in his analysis of *Dasein*, until he finally followed the steep downhill road to collectivism, a development already laid down on the infamous page 384 of *Being and Time*, by filling in the gap with the collectivist notion of *Volks–Dasein* (*Dasein* of the people). Heidegger now answers the question "who are *we?*" with: "the people" (Heidegger [1934] 1998: 59). He now calls the people a collective *Dasein* to which he ascribes the capability to self-responsibility and self-reflection (cf. Heidegger [1933] 1983: 10; see also Thomä 1990: 550). Reading Heidegger's ontology of the people's *Dasein*, one can hardly help getting the impression that Heidegger, in his fixation on individual self-reference, here just replaces the monological self-reference of the single individual with the no less monological self-reference of a collective. In other words, the concept of *Dasein* is simply transferred from the individual to the collective level; Heidegger's individualism thus seems to turn into its opposite within a basically collectivist conception of *Dasein*.

As opposed to this oscillating between individualism and collectivism, inter-intentionality lies beyond this alternative, since the subjects of inter-intentions are neither single individuals nor single collectives, but *individuals in intentional interconnection*. The *subject-we*, as Sartre called it, has no mind of its own. But it would

be wrong to conclude that the subject-we is "nothing but" an aggregate of single individuals. It is one of Sartre's many merits that he saw more clearly than Heidegger himself the inter-intentionalist potential of the concept of *being-with*. In this respect, Sartre is Heidegger's most astute interpreter and his sharpest critic at the same time. Therefore, it might be worth taking the time to have a closer look at Sartre's possible reasons for rejecting a conception of whose potential he was so clearly aware.

Against the theory of the *subject-we* and the pre-intentional, lateral interconnection between individuals, Sartre is determined to insist on the individualist idea that on the basic level, sociality is made of nothing but individuals. Perhaps Sartre's theory of the "pour autrui" is the most consequent of the numerous attempts to base a social ontology on explicit intersubjective experience. In Sartre's theory, no lateral pre-intentional sense of the other, but the explicit experience of others serves as the basic building block of the social. Sartre does not deny the phenomenon of inter-intentionality and the subject-we as such; it's just that he does not believe that it is so important, let alone the basic level of sociality. In defense of individualism, Sartre, in his critique of Heidegger, diminishes the role of the subject-we in any respect he can. Inter-intentional action thus appears, in Sartre, only as a transitory phase, in which individuals temporarily come together only to pursue their respective individual aims. Sartre illustrates this with the example of a mass of commuters jointly using the passages of a subway station. Following Sartre, it is constitutive of joint actions that the single participating individuals "aim at individual goals beyond the presently pursued collective goal" (Sartre [1943] 1991: 497). And in Sartre's example at least, there seems to be some reason to this view indeed. It is true that hardly anyone in a mass of commuters will consider the coordinated use of the subway station an end in itself. Everybody jointly uses the subway, but ultimately goes his or her own individual way. Thus individual action seems to be more fundamental than joint action, because jointness is just a transitory phase in a venture that is ultimately purely individualistic.

Yet Sartre goes one step further in limiting the importance of the subject-we. Even in the communal phases of such actions, he says, the status of the subject-we is only that of a contingent psychological fact that does not in any way reveal an underlying ontological structure. This overt depreciation of the subject-we is so important to Sartre that he repeats it no less than eight times in his chapter on Heidegger's concept of *being-with* (Sartre [1943] 1991: 485, 496, 498–503). In situations such as the one Sartre chooses as an example, some individuals may *experience the feeling* that these are situations of some genuine intentional jointness, but this does not necessarily apply to *all* participants. As Sartre points out, it is by no means necessary that the other participants, too, perceive "us" as a "we", that, in Sartre's words, the others have "an experience that correlates with my experience" (Sartre [1943] 1991: 498). And obviously, Sartre has a point there: in collectives such as the mass of commuters in a subway station, one does not have to have the experience of inter-intentionality, and one does not have to see oneself and the others as a "we", i.e. as a group of pre-intentionally interlinked co-agents. In such situations, we normally tend to interpret ourselves not in terms of some inter-intentional we-subject of a common action, but simply as single individuals, who, in the light of their individual ends, go about their

individual business and coordinate their actions by following the given formal and informal social customs, conventions, rules, and norms.

Thus a distinction between two possible views of social action emerges, bringing us back to the initial question of the relation between intentionality and social normativity. Such situations can be interpreted either as situations of inter-intentional common action, or as situations of individual action in the context of social norms. In the first view, what matters is what we do collectively (i.e. what we do *together*), in the second the focus is on what we do distributively, or severally (i.e. what each of us does for him- or herself). It seems to me that, in light of Heidegger's analysis of *Dasein*, it is apparent that this distinction is of much greater relevance to social ontology than Sartre thinks. It is true that the mode of the individual *One*-self *(Man-selbst)* who conceives of social situations in terms of norm-oriented individual action is the standard everyday-mode of *Dasein*, and only very few people, if any at all, view such situations as situations of joint action. But Sartre is mistaken in concluding that, therefore, the subject-we is merely a psychological phenomenon of no further relevance to social ontology.

Let me explain this *non sequitur* in a direct confrontation with Sartre's own account. Sartre believes that the psychological experience of the subject-we is based on another experience. In order to experience a group of people to which one belongs as the subject of a joint action, one has to experience that group as perceived by a non-member. In other words, the subject-we is based in the object-we, i.e. the group of people is the intentional content of an experience had by an outside observer. Sartre labels this observer 'the third'. Following Sartre, there cannot be the experience of an 'us' without the experience of a 'they' as had by the third. To put it in other words, any form of we-ness and community is based in an experience of being observed in an *I-thou* interaction by a third party ([1943] 1991: 486ff.). Still, following Sartre, it is ultimately the third's view, and not any experience of joint engagement as such, that "ties us together" (Sartre [1943] 1991: 490).

At the basic level, Sartre takes sociality to consist of inter-individual relations of the *I-thou* type (the famous struggle of looks as depicted in *Being and Nothingness*), without any 'we' involved. At this level, there are only individuals and their mutual relations: individuals fighting against each other for their own respective individual possibilities. Here, there is no such thing as a community, but only face-to-face-interaction. Sartre thinks that community comes into play only when some third enters the scene, and makes the individuals and their interrelations the content of his or her own intentional states. The interrelated individuals now see the third observing them in their confrontation, and this experience is what turns them into a 'we'. They experience themselves and the others against whom they are struggling as the 'they' of the third's perception, and this is what turns them into a 'we'. Melting the I and the thou of the original confrontation into a 'we', the third also synthesizes, following Sartre, the participating individual intentions into a *joint intention*. To quote Sartre's own example: it is the synthetic power of the third's view through which my individual intention to beat you and your individual intention to fend off my attack are turned into *our joint intention to fight* (Sartre [1943] 1991: 490).

How plausible is this account? It seems to me that the claim about the fundamental status of the experience of the third clearly has counterintuitive consequences. Note that in Sartre's account it's entirely up to the third who belongs or does not belong to the we-group. The synthetical power of the third's view is strong enough to "stick me down between any other existences", apparently without there being any further criteria (Sartre [1943] 1991: 491). The answer to Heidegger's question, "who are *we*?" is simple: we are the people some third happens to have in his view. The 'we' can be the group composed by some perfect stranger, who just happens to walk down the street ahead of me, and myself, or any other selection from among humanity you might imagine, as long as it only meets the condition that there is some third who somehow "has us in his or her view". Note that this could also apply to the television announcer and me, if only I can see that the third has "us" in his field of vision, and if awareness of the third's perspective need not be mutual.

It seems clear that, while having some plausibility in such examples as children's playing in a room with the third entering the scene as a parent, this is utterly implausible in the previous cases. Here, it is inappropriate to speak of some unification or some "glueing together" of individual intentional states to joint intentionality. The fact that there is some person observing other people, and that (some of) these people are aware of this, does not mean that these people are now justified in using the term 'we' in any other than a purely distributive sense. Obviously, the third's capacity to synthesize individual intentions to joint intentions is not unlimited. There are criteria to be met that go well beyond the mere presence in the third's field of view.

So what are these criteria? I suggest the following answer. The experience of the third's view cannot *create* but only help to *reveal* or *discover* joint intentionality that was already there. A joint intention can be revealed in the third's view only if it was already *latent* in the original situation of action (i.e., before the third's appearance). There is no way my washing the dishes and the television announcer's reading the news can be turned into parts of some joint action intention just by some third's looking through the window, for the simple reason that these activities are not individual contributions to a joint action in the first place. By contrast to this example, there are other cases in which the third's view is indeed important for the interpretation of action. Thus I might interpret my trying to hit you as a purely individual intention, and your fending off my attack as a purely individual act. With the third's appearance, my perception of the situation radically changes. "I fight only if we fight" (Baier 1997: 28) – it now becomes transparent that the individual intentions and actions constitute individual contributions to a joint action. But this is no new fact. Fighting had been a joint intentional activity even before the other entered the scene. What changes with the third's appearance is that this structure becomes explicit. Concentrating on our individual contributions to our joint action, it might well be that we simply loose sight of the fact that our underlying intention, i.e. the intention to fight, is shared – or that we had never been aware of this fact in the first place. Here, the third's view might remind us, or help to bring this intentional structure to light. Thus the third's view does not *constitute* a joint intention. But it may *reveal* a subject-we. If one chooses to retain the idea of the view

of the third, it should not be allotted the task of synthesizing individual intentions, but of breaking an individualizing and over-individualized self-understanding that permeates our everyday life.

Thus Sartre's argument for why the subject-we shouldn't be seen as a fundamental ontological fact about sociality fails to convince on an argumentative level. But this does not mean that Sartre's analysis is worthless. Rather, our discussion of the role of the third points at a solution to our initial problem. An answer to the question of how to fill in the gap in the conception of *Being and Time*, i.e. how to fill the position of authentic *being-with*, and how to distinguish authentic *being-with* from the conventionalism of the *One* seems to emerge. The shift of perspective from an individualized self-understanding of *Dasein* to an awareness of jointness provides the key to a reading of the distinction between authenticity and inauthenticity that is not flawed by an exclusion of sociality from its authentic side. What follows is a re-description of that distinction which uses the results of our discussion of Sartre's account.

In everyday social contexts, we usually act on the basis of an understanding of ourselves as single individuals who go about their individual business, following the guidelines of (and keeping within the limits of) formal and informal social norms and conventions. Thus, in everyday life, we are basically concerned with the conformity or nonconformity of our actions (Heidegger speaks of *Abständigkeit – distantiality*, or, as it is also sometimes translated, *standoffishness*). Traffic regulations might serve as an example to illustrate this. While moving about in the streets, pursuing our own individual aims, we either conventionally stick to those regulations or just strategically avoid being caught violating them, now and then getting annoyed when the regulations appear to be a hindrance or nonsensical in the light of our own or anybody's average individual plans or preferences. This is the inauthentic everyday mode of Dasein's sociality; but what is it that makes it inauthentic? Remember the basic trait of inauthenticity: *Dasein* that is inauthentic lives past its own being (i.e. lives past its possibilities) by being unaware of itself. Inauthenticity is a matter of *Dasein's* covering up its own being by covering up its own possibilities. So what is it that is covered up in the everyday public mode of *Dasein* that I have just described? What remains covered up within this individualized view of the *One*-self is not any authentic atomic self. Quite to the contrary, it is the *communal* character of the underlying situation of which *Dasein* is unaware. In other words: *Dasein* here mistakes its own being for that of an isolated individual, where it is really *joint* or *shared Dasein*.

Superficially it may look as if I did x (e.g., stick to the traffic regulations) because this is how 'one' acts, or should act. The reasons may vary widely; I may act the way I do for fear of sticking out in the crowd, or because of my personal deontologic convictions, or for fear of punishment or other negative consequences. Be that as it may, there is something that I remain unaware of. *Actually*, I do x *as my contribution to our doing y* (e.g. our organizing the traffic), i.e. *because we intend y* (e.g., for the purpose of arranging for a communal life in which everyone can pursue his or her

plans).[12] Thus social norms play an ambivalent role. On the one hand, social norms *facilitate* joint actions simply by standardizing the individual contributive actions required for the common action to take place. Where joint actions are structured by norms, everyone knows which contributions are expected from the participants. *Modern* social norms typically require *uniform* contributive actions, i.e. the same contribution from *everyone's* part. While this greatly facilitates joint action, it also has the following effect. In the self-understanding of *Dasein*, the communal character of action and of the *Dasein* involved here becomes covered up by the individual, norm-oriented *Anyone* (Man-selbst). This individualizing conception of oneself has to be disrupted if *Dasein* is to become aware of its own being and its own possibilities. With this disruption, i.e. the shift from inauthenticity to authenticity, the view of norms radically changes. In the authentic view, norms and conventions are not just restrictions and guiding lines for *Dasein*'s individual actions. Rather, norms are the infrastructure of our *common Dasein*. Norms are the *instruments* with which, with more or less circumspection, *we 'take care' of the Dasein we share*. An authentic view of norms is *sub specie communitatis*, as it were.

Thus it seems that the inauthenticity of the *One* in Heidegger should not be interpreted as standing in contrast to an *individualistic* idea of authenticity that is intrinsically alien to any form of social normativity (let alone to some "authentic *Dasein* of a people"), but in contrast to a *common* or *shared Dasein*, a *Dasein* which is transparent to itself in its common, inter-intentional practices of shared "taking care". As single individuals we can stick to the norms, ignore them, or purposively violate them. But whatever we choose to do, we have already lost our *individual* being, for in situations like these our actions go past our *own individual possibilities*, since they inevitably bear a social meaning that transcends our individual lives. It is only as *common Dasein*, however, that we can *change* or *adapt* norms according to our common aims and ends – and therewith "win our common being" (that was lost in the individualized *One*) in establishing an explicit relation to *those of our possibilities which we do not have as individuals*, *but only as a common or shared Dasein*. The fact that we tend to forget that *Dasein* is not only its individual possibilities, but its *shared* possibilities as well, is what makes our everyday *Dasein* inauthentic.

Let me conclude with some brief remarks on intersubjectivity, an idea that some phenomenological philosophers seem to entrust with the role of the basis of human sociality. I think we should not expect too much from theories of intersubjectivity. From the word 'intersubjectivity' alone it is quite apparent that there is something about this concept which one might either classify as tragical or as comical, depending on one's taste. The word appears to be saying something none of the current theories of intersubjectivity (not to mention those theories which label themselves *intersubjectivist*) actually mean to assert: that the 'inter' can be attributed to the 'subject'. These two elements do not go well together, indeed they exclude each other, at least if we take the subject to be an Husserlian ego which, in its solitary self-reflection reduces all transcendent being to the immanence of its individual

[12] For an analysis of this structure see Sellars (1992: 222); Rosenberg (1980).

consciousness. Intersubjectivity is a *contradictio in adjecto*. The whole point of the subject is that there isn't anything that cannot be reduced to it, while the point of the 'inter' seems to be that there is something that balks at this reduction to the subject. The 'inter' is something ego and alter *share*, which is not to be reduced to either ego or alter *alone*, something *communal*. By contrast to what one might think, intersubjectivity isn't a category from the *philosophy of the subject* (Subjektphilosophie). Rather, it belongs to its *critique*. Intersubjectivity is the title for the search for an alternative to the subject-philosophical *exclusion* of the 'inter' by the 'subject'.

Within this venture, many alternatives have been proposed, among these intersubjectivism, conventionalism (normative pragmatism), and deconstructivism. All these critiques of the *philosophy of the subject* share one basic conviction. In all of these views, intentionality is seen as the epitome of the main problem of the *philosophy of the subject*, i.e. its *monological structure*. By contrast to this perceived monologism, intersubjectivists and pragmatists (such as those mentioned above) claim that intentionality presupposes social normativity, and that, therefore, the analysis of the intentional *givenness* of the world should be replaced by an analysis of the customs and social practices of a community in which the world is disclosed.

Other approaches to intersubjectivity, such as some recent phenomenological, post-structuralist or deconstructivist theories, view the 'inter' as something that goes *beyond* intentionality rather than as one of its presuppositions. Much of the jargon of current French phenomenology is due to this move. Just as in Sartre's account, the 'inter' is conceived of in terms of face-to-face experiences of the other. These theories, too, are critical against earlier theories of intentionality. These critics insist on the fact that the 'inter' comes only at the price of the self-centeredness that is seen as the hallmark of intentionality. The other, it is claimed, eludes any intentional objectivation. He or she is more than what is intentionally "given"; she or he transcends or, as it were, exceeds ego's intentional capacity, thereby defeating the egocentrism of intentionality.

The conjecture exposed in this chapter goes, of course, against the grain of both of these types of theories of intersubjectivity. I find none of these conceptions of the 'inter' particularly helpful, even though both views have a sound core. I do not deny that social practices play an important role for many of our intentional states, and I certainly agree that it is important for our common lives to always be aware of the fact that those with which we share our lives are never just what we believe them to be. But both views completely fail to explain the crucial element of *sharedness* that marks us out as social beings. What we share is neither a set of quasi-*a priori* social practices, nor something that arises from (or is based on) an experience of the alterity of the other that somehow transcends the capacity of intentionality. What we share is primarily a matter of joint attitudes. The sociality of the *disclosedness of the world* is a matter of joint attention, joint intention, shared experience, and shared feelings; in short, it's a matter of joint intentional states. Therefore I propose to see the 'inter' as a feature of intentionality itself, not as one of its presuppositions, or as something going beyond intentionality, as the other theories have it.

It may appear as if the concept of inter-intentionality proposed in this chapter meant nothing else than shrinking the 'inter' to fit the subject, and in fact,

inter-intentionality requires neither an *a priori* of social norms nor any special mysterious alterity of the other. But the theory of inter-intentionality, too, requires a departure from the *philosophy of the subject*, and especially from one of its traits that is uncritically taken over even by its critics. In question is what Annette Baier calls the "Cartesian Brainwash" (Baier 1997: 18): the idea that intentionality is always a matter of the immanence of individual minds alone. Some intentionality is more than that. Some intentionality is genuinely social.

Chapter 10
'*Volksgeist*'

Moritz Lazarus' Social Ontology

Many universities award prizes for outstanding PhD-theses, but only the University of Bern does so in the name of *Moritz Lazarus*. Even in Bern, however, only a small minority will know anything about the person behind that name. The laureates of the Lazarus-prize should at least know who Lazarus was, and why their prize is named after him. Thus, in a short memorandum distributed over the university homepage, some background information is provided. Lazarus, it is explained, co-founded *Völkerpsychologie* (translated as *psychology of nations* hereafter) together with Heyman Steinthal. From 1860 to 1866, the memorandum continues, Lazarus was a full professor at the University of Bern, a time during which he also served as dean of the humanities department, and even as the university's president. Special emphasis is laid on the fact that Lazarus' lectures were attended by an extremely wide audience, attracting even the non-academic public of the city to the university halls. In contrast to these remarks on Lazarus' public success, the comments on the *content* of Lazarus' teaching are much more restrained in tone, to say the least. Even though the memorandum acknowledges Lazarus' efforts to introduce a historical perspective in social psychology, it is stated that his intellectual venture, his psychology of nations, simply "has to be considered a failure".[1]

This is a harsh statement indeed, especially considering the fact that it was Lazarus himself who donated this price. So what is so bad about his psychology of nations that not even the University of Bern, with the best of reasons to do so, can find more positive words? What is the reason for this thoroughly negative view?

It is very likely that the main reason for this negative assessment lies in the *central notion* of Lazarus' psychology of nations. That notion is the *Volksgeist*. According to the most condensed definition that Lazarus gives of his intellectual venture in all of his work, psychology of nations is, very simply, the "science of the *Volksgeist*". Psychology of nations is about giving a psychological description of the "essence" of the *Volksgeist*, and to discover its governing laws (Lazarus [1851–65] 2003: 4, 7, 8).[2]

In our day, however, the very word *Volksgeist*, is something of a red flag. When it comes to that term, even as cautious, balanced and even-tempered an encyclopedia

[1] Cf. www.kommunikation.unibe.ch/communiques/2002/020531lazaruspreis.html (found in January 2005).

[2] After this referred to as GVK.

H.B. Schmid, *Plural Action: Essays in Philosophy and Social Science*, Contributions to Phenomenology 58, © Springer Science+Business Media B.V. 2009

as the *Historisches Wörterbuch der Philosophie* loses its notorious reticence. The *Volksgeist*, the *Historisches Wörterbuch* says, is a thoroughly "compromised notion", and it is with good reason that it is shunned (Grossmann 2001: 1106). This view seems to be almost universally shared. Browsing the relevant literature, the only exception to this rule I know of is a group or movement of followers of the German thinker Rudolf Steiner. Among the Anthroposophists, as they call themselves, the term *Volksgeist* still seems to be a part of their view of the social world, or was at least until well into the 1960s of the past century (Heyer 1990). Apart from this exception, the verdict seems to be unanimous. There is no question that this concept is simply unfit for the analysis of the social world.

§33 The Collective Mind – Past and Present

No wonder the *Volksgeist* has such a bad reputation: even the most cursory look at the history of the concept immediately reveals how deeply the notion is imbued with ideas some of which might appear simply mysterious, but most of which are outright abominable. Indeed, large parts of the history of the *Volksgeist* read like a list of excellent reasons *against* this concept. Here are some examples. Very often, the main purpose of the term *Volksgeist* seems to be to tie a nation's self-determination down to some alleged historical fate or destiny, or to some ready-made boundaries, guiding lines, or some other contingent circumstances which seem external to the process of political self-determination. This is particularly notable in the *German Historical School of Law*, in which the term was widely used. A favorable description of the role of the term in this school can be found in Erich Rothacker's introduction to the *Geisteswissenschaften* – it should be noted, however, that later in his life, Rothacker saw the German *Volksgeist* come to its perfection in Adolf Hitler's rule (Rothacker 1920: 37ff.). But even apart from its association with National Socialism, the concept does not appeal to us. Conceived of in terms of *Volksgeist*, a nation's shared identity is not seen as a matter of the shared intentions and aims of people, and not as a matter of the joint initiatives, shared projects and practices, but as a matter of what people *are*, as a matter of some *given*, of people's adventitious stigmata such as their origin. In the worst case, the unity of the *Volksgeist* is even seen in something like a "racial bond of blood".[3] Even in the case of the pre-Nazi notion of *Volksgeist*, the aim behind the concept is to conceive of social identity in terms of what people *are* instead of in terms of what they *do*. Even here, the concept is accompanied by more or less overt depreciation of both individual autonomy and collective democratic self-determination. Volksgeist seems to be a notion that is both genuinely anti-liberal and anti-democratic. Both of these tendencies are particularly obvious in Othmar Spann's thinking. The aim behind Spann's use of the term is to depart with the principle of self-determination as a guiding theme, and to turn to some "unity" (*Ganzheit*) that is seen as the "nourishing father

[3] An example for this is Larenz (1935); esp. p. 43.

of the human spirit" (Spann 1921: 96–111). In any case, *Volksgeist* stands for anti-individualism and, very often, for a turn away from rationality and the enlightenment as guiding lines, and away from the principles of the French Revolution. In his *Dawn* (also translated as *Daybreak)*, Nietzsche discusses the *Volksgeist* under the significant heading "the Germans' hostility against the enlightenment". The *Volksgeist*, Nietzsche says, is among the "helpmates of the obscuring, quixotic, degenerating mind" (Nietzsche [1881] 1977: 171ff.). Obviously, there is a political agenda behind the concept which is to restore some social substance that according to the proponents of the *Volksgeist* has been eroded by liberalism and democracy.[4]

Sometimes, the concept simply serves to propagate cultural homogenization and the exclusion of otherness.[5] *Volksgeist* stands for the fight against "foreign intruders threatening our own characteristics", and even for the fight against the use of foreign words in the German language (which Otto von Gierke [1915: 24] seems to consider a particularly dangerous transgression against the *Volksgeist)*. Thus the term is directly connected with sheer chauvinism, a role for which it seems to be particularly well suited.[6] *Volksgeist* is a conceptual tool for social *exclusion*, and thus incompatible with any participatory view of membership.

So much for the history of the *Volksgeist*, read as a list of arguments against that concept. It is hardly an exaggeration to say that the *Volksgeist* nicely epitomizes the exact opposite of any normative idea that is near and dear to most of us in our social and political thought. "*Volksgeist*" is the perfect antonym to almost every valid normative political ideal: enlightenment and modernity, liberalism and democracy, the value of individual autonomy and collective self-determination, and the recognition of cultural differences.

Yet the fact that it seems normatively unacceptable is not the only thing that is problematic about the *Volksgeist*. It is also *ontologically dubious*, to say the least. The critics of the *Volksgeist* have expressed their doubts as to whether there could ever *be* such a thing as a *Volksgeist* quite openly, labeling it a mere "phantom". Friedrich Nietzsche expresses his skepticism as follows. It is "dangerous", he says, "to predicate anything of a nation", because this leads to an "illusion of unity" (Nietzsche [1872–1874] 1978: 253). Along similar lines, Georg Jellinek says about the *Volksgeist* that it is "merely a specter" (Jellinek 1914: 153).

This ontological problem, however, is not specific to the *Volksgeist*. Rather, it concerns all the members of its wider semantical family. All conceptions of the collective mind, and of the collective person or subject are confronted with the same skepticism. Some of these ideas might not be as tightly linked to such disastrous normative political ideas as the *Volksgeist*. Yet in an ontological perspective, they are no less dubious. As a look into the relevant literature reveals, the critics of any such

[4] This is explicit in Spann (1921).

[5] Cf. Ludendorff (1933).

[6] For an obvious example see Gierke (1915: 5). Gierke greets the break out of the first world war as a "state in which the Volksgeist finally takes possession of each and every individual soul, and, by coming to life as a great unified 'I', erases any consciousness of the individual I's". Gierke (1915: 29) also claims that the German Volksgeist is much better at that than any other Volksgeist.

ideas have prevailed. Since the end of World War II at the latest, the only role left for the members of the semantical family of the collective mind in social theory and social science is that of a confused notion – and indeed an abominable metaphysical excrescence – from the past, a specter that is effortlessly exorcised by means of a firm commitment to methodological or ontological individualism. Individuals and their mutual relations rather than spooky holistic entities are now seen as the proper object of social science. The only appeal made to such spooky entities is when it comes to justify the individualistic setting of current social theory. Such uses are pervasive in social theory and social science from Max Weber's classical foundation of social science (Weber [1921] 1980) up to the present time. John R. Searle, one of the main protagonists of current social ontology, often mentions the group mind when he comes to characterize the basic traits of his theory. Such conceptions are, Searle claims, "mysterious at best" (Searle 1990: 118), and basically just "perfectly dreadful metaphysical excrescences" (Searle 1998b: 150). To this expression of his deeply felt disgust Searle then adds his commitment to individualism, according to which there cannot be any minds other than those of individuals. There is a routine of declaring the collective mind a terrible idea from the past that is luckily dead and from which current social theory has long parted.

Yet there is another side to the coin. In recent years, there are some new overtones to be heard in the discussion about social ontology and the philosophy of social science. The commitment to one or another form of individualism might still be almost universal, and it is perfectly clear that nobody in the current discussion endorses any of the normative ideas connected to the *Volksgeist* listed above. Yet, in the current struggle for a more adequate understanding of the structure of the social world, there are some ideas around that at first sight at least seem to bear a striking resemblance to some of the other members of the wider family of the collective mind. Leading participants in the contemporary debate, such as Margaret Gilbert, Philip Pettit, or Raimo Tuomela, use terms such as "plural subject" (Gilbert 2000), "groups with minds of their own" (Pettit 2003), or speak, somewhat more cautiously, of "modern counterparts of group minds" (Tuomela 1995: 231). Many philosophers are interested in forms of collective agency that cannot simply be reduced to the agency of the participating individuals. Some philosophers have even started to openly consider the possibility that there is a sense in which personhood might be attributed to collectives in the simple straightforward sense that goes beyond the meaning of corporate personhood in law (e.g., Rovane 1998).

Considering these and other examples, one might ask: is the collective mind really as dead as it seemed? And if it is still alive, or has come back to life in the current debate: what should we make of this fact? Is this yet another effect of some "new collectivism," as diagnosed by Stephen Turner (2004: 386ff.)? And, even more pressingly: what does this recent development mean for the prospects of the *Volksgeist?* Do we have to prepare ourselves for its return, too?

As we have seen, it is not without reason the *Volksgeist* is the most infamous of all conceptions of the collective mind. Yet if so many other members of the family of the collective mind have found their way back into the current debate: what should prevent the return of the *Volksgeist*? Why shouldn't it be expected to be back sometime soon, if so many of its family members already are?

For the reasons mentioned above, this prospect might not seem particularly appealing, and indeed rather frightening. If this is the case, however, it seems that Lazarus' version of the *Volksgeist* would perhaps be the least disagreeable version. It's neither connected to the conservatism of the *Historical School* (Lazarus' views were largely liberal), nor is it in any way in danger to be associated with later Nazi racism in general (Lazarus is above suspicion in that respect) and anti-Semitism in particular (of which Lazarus was an outspoken opponent). So let's have a closer look at Lazarus' concept of *Volksgeist*.

§34 Return of the *Volksgeist?*

Lazarus' effect on later generations does not match the huge success he had with his contemporaries. Not long after his death, Lazarus' work fell into almost total oblivion. It is very remarkable, however, that there is now a new German edition of some selected papers, which appeared in 2003 as volume 551 in Meiner's *Philosophische Bibliothek* series, edited by Klaus Christian Köhnke. In his introduction, the editor does his utmost to make Lazarus' work appealing. In particular, he emphasizes Lazarus' aim to lay a foundation for the scientific study of culture *(Kulturwissenschaft)*. But this venture, too, is directly connected to Lazarus' main conceptual tool. In his view, culture is "objective *Volksgeist*". There is no beating about the bushes here; Lazarus' thinking cannot be advertised without giving that concept a reading that makes it at least half-way presentable. So let's address the core question: what did Lazarus mean by *Volksgeist?*

As emphasized above, Lazarus' concept sticks out from the history of the term in many respects, and it deserves a separate analysis. The following interpretation will be largely based on a passage that has special weight in the context of the whole of Lazarus' work. The passages from which I will quote are taken from the introduction to the first volume of the *Zeitschrift für Völkerpsychologie und Sprachwissenschaft* (which is dated 1860, but appeared in print already in 1859), which Lazarus co-authored with Steinthal.[7] As the introduction to his newly founded journal, this can be considered a programmatic publication indeed. The special importance of this work is also underscored by the fact that Lazarus recycled and further developed large parts of the considerations he presented here later in his life, and he repeated the central passages word for word in a work he published 20 years later.[8] So let's have a closer look at these passages.

Lazarus starts out by saying that *Volksgeist* is "what turns some plurality of individuals into a nation" (EGV: 29). *Volksgeist* is, he says, nothing hovering over and above the heads of the individuals, but rather an "internal bond" (GVK: 12). It becomes immediately clear that Lazarus does not conceive of this bond in terms of

[7] Lazarus and Steinthal (1860), hereafter referred to as EGV. Lazarus co-edited this journal until 1890 (the journal was continued under the title *Zeitschrift des Vereins für Volkskunde*).

[8] Cf. Lazarus (1880: esp. pp. 5–18).

any external, given factor, but in terms of the attitudes of the individuals involved. Lazarus is as explicit on this as one might wish. The bond, he says, is not a matter of a common past, not a matter of any "shared history", not even a matter of such cultural factors as shared religion, shared customs and conventions, shared language, or of the "same type of housing". People can be united without any of these factors. Indeed, for individuals to form a nation it isn't even necessary that they share the same territory or that they have a "common residence" (GVK: 87). To substantiate his claim that these and other more or less external factors are neither sufficient nor necessary for the unity of a nation, Lazarus gives a list of examples. Throughout his career, Lazarus has continually laid emphasis on the fact that the unity of a nation is compatible with vast differences in provenance, cultural origins and influences, religious and linguistic orientation. He does so in his programmatic paper as well as 20 years later in his contribution to the so-called *Berliner Antisemitismusstreit*. In 1879, Lazarus gives a talk at the *Hochschule für die Wissenschaft des Judentums* in Berlin (which was established just a few years earlier based on Lazarus' own initiative). The title of the talk is a question: "What does *national* mean?" In this talk (which was published as a small brochure), Lazarus repeats his earlier thoughts on the ontology of the *Volksgeist*. In this context, the considerations concerning the space for plurality within the unity of the *Volksgeist* appear even more important. Following the same line of reasoning, and continuing to argue against pervasive monolithic constructions of the concept of nation, Lazarus emphasizes that individuals can have more than one nationality (WHN: 17). Indeed there is a further, and more general thesis in the background of this claim, a thesis concerning the relation between the individual and his or her group (nations and other collectives) in general. According to that thesis, multiple group membership is not just a conceptual *possibility*, but plays a constitutive role for the *individuality* of the members of these groups. On the sub-national level at least, it is not just possible, but even *essential* for an individual to be a member of different groups (cf. e.g. GVK: 50). This thesis is known under a label which Lazarus' most famous student Georg Simmel attached to it: "the intersection [or cross-cutting] of social circles" (Simmel [1908] 1983: 305–344).

Based on these few characteristics, it seems safe to say that Lazarus' *Volksgeist* is no monolithic or even uniformist idea; and quite obviously, it in not a matter of contingent givens or of any adventitious stigmata either. In a sense, even the talk of the "being" or "ontology" of the nation is misleading. According to Lazarus, there *is* no nation. The unity of the nation is always a *process*, and not a matter of any substance (EGV: 27; WHN: 13). Lazarus adds that this is the reason why he chooses the term *Volksgeist* rather than *Volksseele* (the *soul* of the nation).[9] Moreover, the

[9] "Wenngleich nun aber auch eine Substanz des Volksgeistes, eine substanzielle Seele desselben nicht erfordert wird, um die Gesetze seiner Tätigkeit zu begreifen, so müssen wir doch jedenfalls den Begriff des Subjects als einer bestimmten Einheit feststellen, um von ihm etwas prädiciren zu können" (GVK: 11). This brings Lazarus into sharp contrast with Wilhelm Wundt. Wundt claims the Volksseele to be the object of his Völkerpsychologie, because by contrast to the spirit or mind *(Geist)*, the soul is embodied (especially in cultural artefacts; cf. Wundt [1900: 7]).

process in question – which in a sense *is* the nation – is constituted from the *participants' perspective*, not from any observer's point of view. This is an insight that is so important to Lazarus that he repeats it on several occasions: "[the unity of a nation] is a mental product of its members" (EGV: 36; GVK: 89). The constitution of the nation is due to a kind of a reflexive self-categorization, in which the individuals take themselves to be members of the nation.

> What makes a nation a nation lies in [...] the subjective view of the members of that nation, who *see themselves* as a nation. The concept of a nation is predicated on the nation's members' subjective view of themselves. (EGV: 34–35; GVK: 88)

Thus it is clear that the unity of a nation is not a matter of an irreducible collective substance. In Lazarus' thinking, *Volksgeist* is simply the title for the process of the individual members' subjective self-categorization. To use the expression proposed by Benedict Anderson in his influential analysis of the concept of the nation, it seems that, in Lazarus' view, nations are simply "imagined communities". Indeed, the similarity between Lazarus' view and Anderson's is striking. Anderson's central concept of "imagination", just as Lazarus' *Volksgeist*, is ultimately a matter of self-reflection and self-interpretation.[10]

But why, then, choose the term *Volksgeist*, which seems to ascribe the mind or spirit in question to the nation rather than to its individual members? An influential and particularly piercing critique of the assumption of the *Volksgeist* was put forward by Wilhelm Dilthey, a long-time friend of Lazarus', who in 1866 aspired to Lazarus' succession at the University of Berne,[11] but then accepted a call to the University of Basel that same year. At the core of Dilthey's critique of the *Volksgeist*, which was not directed against Lazarus' version, is the claim that, for there to be mind, there has to be some self-awareness, but since there cannot be anything like *self-awareness* on the collective level, it does not make sense to credit collectives with any form of mind of their own (Dilthey 1923: 31). Superficially, Lazarus seems to disagree with Dilthey's view. For him, collective self-awareness is no less than the "core" of the *Volksgeist* (GVK: 91), and thus the point of reference for the whole of *Kulturwissenschaft*. Lazarus explicitly states that collective self-awareness is the "most essential element for the definition of a nation" (GVK: 83): "just like each individual, each nation has self-awareness of its own, through which it becomes a particular nation, just like the former becomes a particular person" (GVK: 89). Yet Lazarus makes it immediately clear that the only place for this seemingly mysterious "self-awareness of the nation" is exclusively "in the mind [*Gemüth*] of the individuals" (loc. cit.). The self-reference of the collective is not a matter of some mind hovering over and above the single individuals, but the very act of self-categorization of the participating individuals as encountered above.

In other words: the mysterious "self-awareness of the nation", again, is simply a matter of the attitudes and perspectives of the participating individuals. If we follow

[10] Anderson quotes approvingly from Hugh Seton-Watson: "All that I can find to say is that a nation exists when a significant number of people in a community consider themselves to form a nation" (Anderson 1991: 6).

[11] Cf. Dilthey's letter to Lazarus in Lazarus and Steinthal (1986: 786).

this line of thought further (though there are some passages in Lazarus' work that do not seem to fit into this view),[12] it seems that, behind the specter of the *Volksgeist*, there is an ontology of the collective that is solidly individualistic. Lazarus' view of the *Volksgeist* is fully compatible with Max Weber's later conception that has been of so much influence on later thought. Just as in Weber's methodology, collectives come into play only as parts of the content of the intentional attitudes of the individuals. Collectives exist only because individuals (in Lazarus' version: its individual members) *think* that they exist, or *take* them to exist.

Though the term *Volksgeist* has since fallen into disrepute, the conception behind the term has been of tremendous influence and success, even if it is very rarely ascribed to Lazarus. In the current literature, this conception is usually credited to Lazarus' student Georg Simmel. In his famous '*Digression on the Question: How is Society Possible*?' Simmel derives the unity of social groups from the consciousness of unity of its individual members (Simmel [1908] 1983: 21ff.). *Via* Simmel, this idea has become part of the classical canon of ideas in social theory and social science. Through Simmel, this idea still influences the current debate. Margaret Gilbert, whose ontology of social facts is among the most discussed in the current debate, directly relies on Simmel, calling her view of the ontology of groups "Simmelian" (Gilbert 1989: 146–246). It should not be forgotten that behind Simmel's view is the concept of *Volksgeist* as analyzed by Simmel's teacher Moritz Lazarus.[13]

If some of the core ideas of Gilbert's Plural Subject Theory follow in direct succession from the *Volksgeist*, this does not compromise the theory, because the kind of *Volksgeist* that is at issue here seems to be entirely free of any of the horrors that are usually associated with this concept. By contrast to other versions of the *Volksgeist*, Lazarus' version is of the pluralistic, subjectivist, and individualist kind. It seems that here, *Volksgeist* is by no means a matter of some pre-determined fate or destiny. It is not in conflict with the modern individualistic conviction that the substance of the social – if there is any – is a matter of the individuals and their mutual relations. *Volksgeist* does not limit individual autonomy, since it is *based* on the subjective attitudes of the individuals. Also, according to this notion, nobody seems to be *excluded* from the nation but those who exclude *themselves* by not taking

[12] Thus Eduard von Hartmann, in his analysis of the "Essence of the Whole", quotes a passage (of which he approves) from Lazarus' work, according to which, in modern terminology, the Volksgeist should be seen as a matter of the unintended consequences of individual action, rather than as a matter of the intentional content of individual attitudes: "each one does, what he does, immediately only for himself. Nevertheless, all individuals form a unity through their labour, even unknowingly and unwillingly. This unity consists of real, concrete, and often influential causal relations, that are objectively revealed in the actions of the individuals, only that they elude the individual's awareness, his intentions, and goals" (Lazarus quoted in: von Hartmann [1871: 28ff.]). Here, the *Volksgeist* is a matter of the invisible hand rather than the content of individual self-categorization. The two readings do not seem to square, and I know of no remarks from Lazarus' side of how to relate the two. For reasons quoted above, I think that the self-categorization view of *Volksgeist* should be considered the more important reading.

[13] On Lazarus' influence on Simmel see Köhnke (1996), esp. pp. 386ff. For a somewhat more cautious assessment of that influence see Canto i Mila (2002).

themselves to be its members. Thus there seems to be nothing unduly exclusive or collectivist anymore about the *Volksgeist*. Rather, the *Volksgeist* here seems to take on a completely different, a participatory, indeed a liberal-communitarian hue. Paraphrasing Lazarus, one could say that the *Volksgeist* is about cultivating our shared self-understanding. This is why the *Volksgeist* is never just there, never just given or fixed, nothing pre-determined, but always to be created and to be maintained within the process of continuing reflexive re-interpretation. Thus it becomes obvious that Lazarus' *Volksgeist* contains a view of *collective identity* that is widespread in the current literature. According to this view, collective identity is the "capacity of individuals to declare themselves the community which they are and which they want to be, always anew and afresh."[14] In other words: the *Volksgeist* is but another label for what one could call the *reflexive* account of collective identity.[15] It comes as no surprise that the editor of the new edition of Lazarus' collected papers lays special emphasis on these traits of Lazarus' psychology of nations (GVK: ix–xi). In the light of these considerations, Lazarus' conception of the *Volksgeist* seems surprisingly acceptable, indeed even appealing.

In the remainder of this chapter, I would like to question this positive assessment of Lazarus' *Volksgeist*. In the next section, I will argue against the reflexive view of collective identity, and I will show how the study of Lazarus' work can help us to understand the flaws and limitations of this view. This is important because this view of social identity predominates in current social theory. If there is a problem with Lazarus' *Volksgeist*, this is not *in spite* but *because* of its appeal to current views of collective identity.

§35 Lazarus' *Volksgeist*: Some Problems

If the *Volksgeist* is compatible with the reflexive account of collective identity that is so frequent in current literature, and that seems to be free of any fatalism, collectivism and exclusivity, does that mean that this concept (if not the word) is without problems? I will argue that the answer should be in the negative. As I will try to show, Lazarus' theory of the *Volksgeist* is not just surprisingly modern. It can also help us to see some *problematic consequences of the modern reflexive conception of social identity* that are overlooked in much of the current debate. If Lazarus' notion of *Volksgeist* is surprisingly modern, this should not simply be seen as an

[14] Cf. Tietz (2002: 77, 150, 207).

[15] Following are some current examples for this reflexive line of theorizing collective identity. A particularly prominent example is Tamir (1996), esp. pp. 176–177. Tamir's emphasis, too, is on the fact that identity is changeable. Tamir, too, bases identity on a reflexive self-reference of the individual members: "The quest for identity (. . .) is marked by self-reflection, by the readiness of individuals to make radical changes in the way they perceive themselves" (loc. cit.). Another example is "Social Identity Theory": "Social identity is self-conception as a group member" (Abrams and Hogg 1990: 2ff.). Another example is Matthiesen (2003). For a critique of these reflexive accounts of social identity cf. Schmid (2005c, 2005d).

advantage, but also as one of its problems. Studying Lazarus' *Volksgeist* thus can teach us a lesson. This lesson I believe to be so important that it should be considered one of the foremost reasons why studying Lazarus' work can still be rewarding.

In a first step, Lazarus makes clear some of the logical consequences of his theory of *Volksgeist*. The point of departure is this. If a nation is *constituted* by the respective individuals' taking themselves to be its members, this self-categorization has to be considered *infallible*. As Lazarus says, the social self-consciousness can "never be mistaken" (EGV: 36; GVK: 90). If individuals *take themselves* to be a nation, they simply *are* a nation, because the *being* of a nation, conceived of in terms of a process, is that of these individual attitudes. If this is true, this has further consequences for the analysis of collectives in social science. The main consequence is that the analysis has to follow individual self-categorization. Lazarus finds the following words for this:

> As far as plants and animals are concerned, it is the *scientist's task* to categorize them according to the objective features of the species; by contrast to this, we have to *ask* human beings, to which nation they *count themselves*. [...] We have to elucidate the subjective definitions that nations tacitly (*implicite*) give of themselves. (EGV: 35; GVK: 88; WHN: 13)

In contrast to the natural sciences, epistemic authority in the social sciences lies with the *objects of analysis*, and not with the scientist. For all the democratic and participatory flair of this account, however, there is also a problem to be considered here. At least on a conceptual level, it seems quite important to see that reflexive self-image and lived community are two different things. Why should the question whether or not the most relevant collective entities are those of which the participating individuals are aware be considered a settled matter *a priori*? Why not leave open the *conceptual possibility* that our reflexive self-categorization misses those forms of communal life that are most essential to our societies? Could it not be the case that we (as a community, or as a nation) have long ceased to *be* the community, or nation, which we still *take ourselves* to be? Could it not be the case that unbeknownst to ourselves, we have become a different community, or nation, i.e. that our communal self-awareness misses our actual communal being?

Be that as it may, the problem is that there is simply no conceptual room for such questions if we follow the reflexive theory of collective identity. And Lazarus, with his declaration of infallibility of the collective self-awareness, has made this consequence clear. As we shall see, the fact that among all types of collectives, Lazarus' theory is about *nations*, makes this consequence particularly difficult to accept. The reason why Lazarus' concentrates on the *Volksgeist* rather than on the mind or spirit of any other collective is that he thinks that the nation is, as he says, the "most essential" form of social life (EGV: 5).[16] On one occasion, Lazarus captures this thought in an admirably ambiguous statement: "the form of the common life of humanity is

[16] Later on, Wilhelm Wundt gives a similar answer to the question why social psychology should be psychology of nations, rather than psychology of any other kind of collective: "the nation is the most important among the circles of life (Lebenskreise) from which the products of mental life emerge" (Wundt 1900: 3).

its being divided into nations" (GVK: 52). Yet in view of rapid social change and globalization we have to consider the above questions in all earnestness. Could it not be the case that, in a given historical situation (especially ours), the concept of a nation still predominates in our self-categorization, while having long lost some or most of its relevance on the level of social reality? What if, unbeknownst to us, social reality has become *post-national*, to quote one of Jürgen Habermas' titles? In my view, we should leave room for this, if only in terms of a *conceptual* possibility. But this presupposes what Lazarus denies: in modern terms, it presupposes that it is *conceptually possible* to distinguish between social *self-description* and actual *social structure*. It is well possible that our communal or collective self-consciousness is mistaken. Identifying collectivity with awareness of collectivity, as is done in Lazarus' conception and so many more current theories of collective identity, means to short-circuit things that should be carefully kept separate. It means reducing social theory to a kind of hermeneutics of self-categorization, and this means barring the prospects of a *critical role* for social theory (and theory of society). Social theory is not simply hermeneutics. It is *critique*, too. And in a critical perspective, social theory might not only teach us that our self-categorization and the actual structure of our societies might diverge. More than that, it might teach us that this divergence is of systematic character. An example of this view can be found in the work of the German sociologist Niklas Luhmann. Luhmann's suspicion that motivates most of the huge body of his work is that the structure of our societies has greatly changed, leaving the semantics of our societal self-description far behind. Luhmann argues that the main problem in social sciences is this very gulf which has opened up between social structure on the one hand and the semantics of societal self-description on the other. Our thinking about the social, that still follows the old categories, has simply lost contact with actual social reality that is so rapidly evolving.[17]

Even if the rift between self-understanding and social structure is not quite as deep as Luhmann conjectures, or even if there is no *actual* rift at all, it is not obvious why this should be ruled out *a priori*, i.e. on conceptual grounds, as is done in the reflexive approach. Lazarus and his followers may be right in claiming that social identities cannot be determined purely from the outside perspective. But it should also be considered that the participant's perspective might not always be right, and the ultimate epistemic authority either. Indeed in view of the darker chapters of the history of the *Volksgeist*, one might even think that the more the spirit of a group is conjured up within a group, the less it is real as a living community. If this is true, the connection between the community and the awareness of community might be exactly the reverse. The more insistently and decidedly we try to see ourselves as members of a community, the louder our appeals to the community spirit, the less likely it becomes that there actually *is* a community.[18]

Contrary to what Lazarus thinks, collective self-consciousness is not infallible. Sometimes it is not really *us*, the "we", whom we are conscious of, because there is

[17] Luhmann's systems theory sets itself the task to close this gulf. Cf. Schmid (2000: 124ff.).

[18] This is a lesson taught by Martin Heidegger, who has learned it the hard way himself (see Heidegger [1938/39] 1997: 329).

no such thing as a "we". Social self-consciousness does not *constitute* a community. Rather, it *presupposes* community. The existence of a community is what makes our belief that we are its members true, and not the other way around. If individuals see themselves as members of a team, they assume that this assumption is justified to the degree that they really *are* a team. Thus the very structure of collective self-consciousness defies any attempt to use it as the base of social ontology. Self-categorization presupposes membership, and not the other way around. Lazarus' approach, just as those current approaches to social identity that follow the reflexive line, are simply circular. If some form of intentional states is what constitutes collectives, these collectives cannot be presupposed in the intentional *content* of the mental states in question.

Therefore it seems that we have reached an impasse. What makes the encounter with Lazarus' theory so rewarding, however, is that he not only spells out the implications of a reflexive account of social identity. He is even aware of the consequences I just mentioned. Lazarus quickly adds to his definition of a nation as a "crowd of human beings who take themselves to be a nation" the observation that this definition contains a "logical error" (GVK: 88, EGV: 35). Collectives cannot be born out of their own heads, as it were. If collective self-awareness is consciousness *of* the community *by* its individual members, the existence of the collective is presupposed in collective self-awareness. By contrast to some current philosophers and social theorists who gladly accept any charges of circularity and inconsistency, Lazarus is far too serious a thinker just to let the matter rest at that. So how does he resolve the problem? Lazarus confronts it in an attempt to break this circularity. The way he does so, however, is rather telling. Lazarus now claims that it is not really the *collective itself* that is the content of the intentional state that constitutes a collective. Rather, the consciousness in question is always based "on such objective factors such as origin, language, etc." (GVK: 89ff.). It is true that Lazarus quickly adds that these objective factors are not what is most important about the *Volksgeist*, which continues to be the "subjective and free act of self-conception as a unit and as *one* nation" (loc. cit.). But still, the consequences of Lazarus' breaking the circle by appealing to objective factors are grave, to say the least. Now, there does not seem to be much leeway left for the subjective and free act of self-categorization anymore. The subjective act of self-categorization is bound to grasp those objective factors that are already there, pre-consciously and pre-politically, as it were. It seems that with this move, the *Volksgeist* reveals its exclusive face again. It is not a matter of spontaneous self-invention, of shared imagination anymore, but tied to those very objective factors which Lazarus rejected so vehemently in his original definition of the nation. It seems that all claims to the contrary notwithstanding, the *Volksgeist* is ultimately a matter of fate rather than a matter of initiative, a matter of origin instead of a matter of spontaneous, free, and sometimes border-bridging joint practices, projects, and initiatives.

The first book which Lazarus published – it came out in 1850 – was a defense of German national unity under Prussian hegemony. In a sense, this beginning is significant for Lazarus' published work in general, and his theory of the *Volksgeist* in particular. Lazarus' theory of the *Volksgeist* is deeply connected with the question

of German national unity, and this is its historical place. The question is: beyond history, is there anything this conception of the *Volksgeist* could teach us today? Is there any possible use connected to this concept for the denizens of a social world which Jürgen Habermas (1998) has labeled *post-national*? Or should we simply leave Lazarus' work to the history of ideas? I think the above considerations show that there is a point to be made in favor of a continuing or renewed dialogue with Lazarus. There is indeed something that can be learned from Lazarus' theory of the *Volksgeist* that transcends history. But this lesson can be learned only if we are prepared to see what Lazarus' ultimate failure teaches us about the problems of *our own conception* of social identity first. And this is a lesson many might not like to learn.

It is obvious that, in spite of all appeals to a liberal, participatory conception, some traits of Lazarus' *Volksgeist* point towards the darker chapters in the history of that concept. Here, too, the *Volksgeist* reveals its exclusivist features. Thus the editor of the multi-volume Lazarus-Steinthal correspondence even judges that there is a certain amount of chauvinism in Lazarus' thinking on the *Volksgeist* that all but matches the chauvinism of his opponent in the *Berliner Antisemitismusstreit*, Heinrich von Treitschke (cf. Belke 1971: lxxi). In the sparse literature on the topic, some form of negative assessment is almost universal, even though, in view of the given references, some of the statements seem overly harsh.[19] Independently of how much weight these tendencies have in Lazarus' thought, however, it seems clear that the most important result of the above reading of Lazarus' theory of the *Volksgeist* is the following. If Lazarus' conception of the *Volksgeist* is problematic right down to its core, this is not because it is incompatible with any form of individualism. As far as the passages on which I have based my reading are concerned at least, Lazarus' conception is thoroughly individualistic. The problem of this conception does not lie in any of the tendencies usually associated with the *Volksgeist*. Lazarus' *Volksgeist* does not displace individual agency, it does not bypass the theoretical, practical and affective attitudes of the participating individuals as the base of social ontology. Rather, the problem is that Lazarus narrows the relevant kinds of attitudes to *reflexive forms* of consciousness. This is epitomized in Lazarus' claim that social unities exist insofar and only insofar as its members see themselves as its members. This conception is circular, and Lazarus believes that this circularity can only be avoided by basing the *Volksgeist* on objective factors, which leads back into the problems usually associated with the term.[20] If any of the ideas associated with Lazarus' *Volksgeist* should turn out to be of use to us at all, it is clear that this has to come at the price of a radical reconceptualization that involves loosening the tight ideas of unity that are at play in Lazarus' theory.

Let me briefly point out how this could be done. I think that Lazarus is right in emphasizing the subjective over the objective, and in diagnosing the logical

[19] Cf. e.g. Schneider (1990: 68ff.).

[20] Indeed, in a certain sense, Lazarus' individualism rather than any collectivism is at the heart of the problems of his account. Thus Wilhelm Wundt criticized Lazarus' concept of the Volksgeist as a "projection of the individual mind on the larger scale" (Wundt 1900: 19).

problems associated with reflexive notions of social identity. The problem is that he takes the wrong turn to avoid the circle, taking a path that leads him back towards the exclusive conceptions of social unity which he originally tried to avoid. If Lazarus is right in emphasizing the fundamental role of the attitudes of the participating individuals, his mistake is to concentrate on *reflexive* attitudes, i.e. those individual attitudes that involve some self-categorization. There are pre-reflexive forms of intentional attitudes that are relevant to our communal lives, of which we may not be reflexively aware. Indeed it may well be that our self-understanding misses the kind and content of our pre-reflexive attitudes. Sometimes, what people think, do, and feel, is not exactly what they take themselves to be thinking, doing, and feeling. And many of these pre-reflexive attitudes are *shared*.[21] These, and not any reflexive forms of consciousness, should be considered the base of social ontology.

If we follow this line of thought further, many insights which can be learned from Lazarus become important. Among the lessons to be learned is Lazarus' insistence on the role of plurality of membership, and its significance for our individuality. In this context, Lazarus develops a critique of a distorted view of social unity, that was as wide-spread in his day as it is now, and that is even part of the very word "collective". As Lazarus remarks, standard analyses of social unities usually start out with the assumption of fully developed individuals with all their psychological qualities, their personality, their beliefs and preferences, and then go on to conceive of collectives as something that is *composed* of these individuals. Here is what Lazarus has to say about this line of analysis:

> While appearing just to express the facts, this view implies a tremendous mistake: those qualities and relations of the mental life and that content of the inner being are not inherent to the individuals, conceived of as single beings. Only in social contexts, i.e. only insofar their lives are shared [...] do these individuals acquire and possess the content even of their individual lives. To think of humans in abstraction of their sociality, to conceive of them as bare single beings [...] would be a mere fiction that contradicts all facts. (GVK: 81ff.)

This critique anticipates much of what is now discussed under Philip Pettit's label *Common Mind* (cf. Pettit 1996: esp. 111ff.). Lazarus criticizes an *atomistic* view, and argues for a *holistic* understanding of the relation between the individuals and collectives. Yet clearly, his rejection of a view that always takes the individual to be ready-made, as it were, does not as such entail the slightest denunciation of individual autonomy. Quite to the contrary, a more holistic view is the *condition* of positing the autonomous individual as the "purpose of community", as Lazarus says explicitly (GVK: 113).

Thus we might conclude that Lazarus' *Volksgeist* has very different, indeed incompatible traits. Among the more somber features is the vehemency with which Lazarus, 2 years after the defeat of the revolution of 1848, insisted on the role of German unity, which he seemed to value much more highly than democracy and

[21] For a more detailed account cf. Schmid (2005c).

individual liberty.[22] Under the title *Volksgeist*, national unity is once again preferred to a republican constitution. Yet there is a counter-tendency even to that to be found in Lazarus' work. Lazarus' decision to accept a call to a Swiss university was not only motivated in the sad fact that, as a Jew, he had no prospect of ever becoming a full professor in Germany. As Lazarus recounts in his autobiography, there was a positive motivation behind that move, too. He went to Switzerland – which under the influence of liberal democrats only recently (1848) had adopted a new federalist constitution – "to study the republican life of a nation".[23]

[22] Cf. Lazarus (1850: 50); see also Belke (1971: xlviii).

[23] Lazarus, Nahida (1910: 99). Nahida Lazarus suggests that her husband modeled his view of the relation between individual and community on the Swiss form of government (cf. ibid.: 54ff.).

Chapter 11
Evolution by Imitation

Gabriel Tarde and the Limits of Memetics

When Sigmund Freud counted Darwinism among the most severe blows which human self-love has suffered at the hand of science, he was only referring to Darwin's insight into man's "descent from the animal kingdom" and his "ineradicable animal nature". Had Freud had any apprehension of what else Darwinism had in store for us, he might not have called his own central insight (i.e. the discovery of the role of the subconscious) an even "more wounding blow" (Freud 1957: 84–85). Our animal (and, for that matter, vegetal) kinsfolk, as well as the dominant role of our subconscious, seem rather easy to put up with as compared to the genetic neo-Darwinian image of ourselves. The ultimate blow to our self-love is this. Whereas Darwin himself kindly left us with the belief that the 'struggle for existence' was all about *our* existence (i.e. the existence of the kind of beings that we are, and the kind of life that *we* live – as individuals, as groups and as a species), this picture has radically changed with one influential interpretation of the integration of genetics into Darwinism. According to this view of Neo-Darwinism, it is not *we* – the individuals, groups or the species – there in the spotlight on the stage of the evolutionary drama called 'the survival of the fittest' anymore. It's the genes. The evolutionary story which has been put forth by Richard Dawkins in his famous Selfish *Gene* (1976) is told from the 'gene's-eye perspective', which differs from the perspective of the kind of beings that we are. Evolution, it is claimed, is all about the replication of genes. Whereas the genes are the actors on the stage, we are nothing but more or less contingent accessories. In Richard Dawkins's words, the living bodies and conscious minds that we are have no more importance in the evolutionary story than as the 'survival machines' that some sets of more or less cooperative genes have built themselves in order to provide for their own survival, or replication. "They are the replicators and we are their survival machines. When we have served our purpose we are cast aside. But the genes are denizens of geological time: genes are forever" (Dawkins 1976: 37). In other words, we are "robot vehicles blindly programmed to preserve the selfish molecules known as genes" (Dawkins 1976: ix).

H.B. Schmid, *Plural Action: Essays in Philosophy and Social Science*, Contributions to Phenomenology 58,

§36 The Meme's Eye View

Not only are 'we' replaced by 'them' as the agents on stage; upon closer look, it becomes apparent that, at least at this basic level of the theory, any sort of agency is removed from the picture. The sense in which the genes become the 'subjects' or 'agents' in this story is metaphorical. Genes are not literally 'selfish', nor are they 'programming' anything; 'Darwinism' is the title of the theory that explains why it *looks as if* there was some selfish scheming going on from the side of the genes, even though there are really nothing but 'blind' natural forces, and no design or purpose (in the proper sense of the word) involved in the process.[1] Thus the intentionalist vocabulary (the one that includes words such as 'selfishness' or 'programming') in this neo-Darwinian story is only shorthand for a more complete non-intentionalist description.[2]

As a matter of course, the gene's-eye view of the evolutionary story is not uncontested. Among the most oft-quoted problems of this view is the following. Dawkins claim that the rest of the living organism is simply a "survival machine" for the genes seems to be at odds with the fact that, according to a very plausible view, genes are *defined* by their functional role within that organism, which leads to a holistic understanding of genes. In contrast to this, Dawkins claims the genes to be ontologically prior. The argument for his view is that the DNA must have been there even before there were cells and organisms, and that we cannot explain the genesis of cells without understanding the mechanisms of replication of the DNA. This does seem right, but at the same time, Dawkins' DNA-fundamentalist view of genetics (as I shall call it in the following) has problems answering the question concerning the unity of the gene, i.e. the questions of which sequences of the DNA should be considered as a gene. By contrast to the DNA-fundamentalist view, the functionalist view seems to have considerably less difficulty answering this question.

We do not have to delve any deeper into the question concerning the ontology of the genes here, even though we will run into similar questions below. At this point, our question is a different one. Granting that the *gene's-eye perspective* on biological evolution is right: why should social scientists bother? However severe this blow to our self-understanding, it might seem that the damage can be restricted to our self-image as *biological* creatures, and kept away from our role in society and culture. Our biology might be left to 'them'; but surely, one might think, society and culture are still up to 'us', i.e. a matter of *our* ways of existence. Much to the reassurance of our self-love, most neo-Darwinian attempts to break into the sphere of the social and cultural sciences have proven to be rather limited in range. For all of its fundamental insights, classical sociobiological explanations do not seem to have gone much beyond the level of the analysis of some basic behavioural dispositions

[1] In Daniel C. Dennett's words, Darwin's theory is "a scheme for creating design out of chaos without the aid of the mind" (Dennett 1995: 50).

[2] A gene is said to 'aim at' x when it is selected for the fact that under circumstances y it will cause x. In Dawkins's sense, a gene is ultimately 'selfish' (or 'aims at replication') because of (and insofar as) the fact that, under suitable circumstances, copies of that gene are generated serves as the reason for why it is still around.

so far. And there is reason not to expect much more from this side. There seems to be a systematic barrier to this research program. Many societies and cultures on the planet have undergone fundamental changes within centuries or decades, sometimes even within years. All those changes and new phenomena that have appeared on the cultural scene can hardly be explained in terms of genetic evolution, for at the level of the human genome, hardly anything will have changed within this short time span. The huge difference in speed between genetic evolution and sociocultural developments seems to thwart any attempt to gain substantial explanations of actual social and cultural phenomena by going back to the level of genetic evolution.

In spite of this 'failure' of classical sociobiology, there is no reason for human narcissism to feel safe from the Darwinian blow to self-love in the social and cultural sciences. At the end of his fascinating book on the selfish gene, Dawkins has sketched a neo-Darwinian perspective for the social and cultural sciences, which avoids the classical socio-biological short-circuit between the cultural and the genetic level. It is here, where the memetic project is initiated. The core idea is "to throw out the gene as the sole basis of our ideas of evolution" (Dawkins 1976: 205), and to postulate a second and much faster evolution, a cultural evolution that has started only with the development of human consciousness. Like any evolution, this second one is about the "differential survival of replicating entities" (Dawkins 1976: 206). In this case, however, the 'replicating entities' are not genes, but 'memes', i.e. units of culture such as the ones mentioned in Dawkins's famous list of examples: "tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches" (Dawkins 1976: 206).[3] The close analogy (or kinship) between memetics and the theory of genetic evolution is evident from the word "meme" alone. As Dawkins says explicitly, it is because it "sounds a bit like 'gene'" that he has created the neologism "meme" (Dawkins 1976: 206). Thus it seems that some of the more cautious memeticists' worries notwithstanding,[4] the project to build the theory of cultural evolution on the model of genetic evolution lies at the very heart of the memetic program. Even though not all memeticists agree on whether or not there is an equivalent of the distinction between genotype and phenotype in the cultural sphere, the analogy between memes and genes (and with this the analogy between Memetics and Neo-Darwinism) is not marginal, but a premises of the entire memetic venture.[5] All memeticists agree that memes are replicators just like genes. As such, the evolutionary triad of replication, variation and selection applies to memes just as it does to genes. Whereas genes replicate by *inheritance*, memes replicate by *imitation*; indeed this way of replication is taken to be so essential to memes, that sometimes, memes are simply defined as "units of imitation" (Dawkins 1976: 206).

---

[3] Daniel C. Dennett characterizes memes as "ideas", and he gives the following list of examples: "arch, wheel, wearing clothes, vendetta, right triangle, alphabet, calendar, the *Odyssey*, calculus, chess, perspective drawing, evolution by natural selection, impressionism, 'Greensleves', deconstructionism" (Dennett 1995: 344).

[4] For a more cautious view of the relation between genetics and memetics cf. Blackmore (1999: 66).

[5] Sometimes, the memetic project is criticized on the grounds that the evolution of memes is Lamarckian rather than Darwinian. For a convincing refutation of this view cf. Dawkins's remarks in the foreword to Blackmore (1999).

Whenever somebody takes over something from somebody else by means of imitation, a meme is replicated. And just as genes are subject to mutation, variation comes into play on the memetic level, too, because imitations are rarely perfect. Urban legends (another often-quoted example for memes) are likely to undergo changes along their long way through society. Besides replication and variation, there is also a great deal of *selection* going on in the cultural sphere. Under the current conditions, at least, memes depend on the human mind for their replication. And as this is a rather scarce resource, memes are in fierce competition with each other. The space for memes is limited. Of all the stories we tell each other, only very few will ever make it to the level of an urban legend. Most are bound to fall into oblivion shortly after they are told. "Tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches": all of these are thus in a kind of selective "struggle for existence" (Dawkins 1976: 206).

After Dawkins seminal contribution in the last chapter of his 1976 book, the memetic program has gained considerable momentum. Besides Daniel Dennett (1991, 1995), there are some more names to mention. It is perhaps no coincidence that many memeticist originate from the margins between academia and the wider public. Douglas Hofstadter (1985) has been instrumental in spreading the idea in contributions to the *Scientific American*. In 1996, Richard Brodie, who had already gained considerable reputation as one of the minds behind *Microsoft Word*, published his book *Viruses of the Mind*, in which he focuses on the central memeticist idea that truth is but one of the factors that might explain why beliefs spread through a population. The same year saw the publication of *Thought Contagion: How Belief Spreads Through Society* by Aaron Lynch. Lynch claimed to have developed the core thoughts of his book independently of the memetic movement. In 1997, the first issue of the *Journal of Memetics – Evolutionary Models of Information Transmission* appeared, and Susan Blackmore published her important book *The Meme Machine*.

It is characteristic of Blackmore's account as well as of those of the other contributors to the debate that Dawkins' *perspective shift* is transferred from the biological to the cultural level. The general thrust of the memetic program parallels that of the DNA-fundamentalist neo-Darwinian approach. Just as the latter wipes us from the centre of the stage of biological evolution, the memetic program displaces us from our high seats as the authors and creators of our 'cultural products'. In this view, we are not the ones in charge in the sociocultural sphere: it's them, the memes. Culture is not to be conceived of in terms of what *we* think, feel or decide to do anymore. In the memetic perspective, it is simply a matter of the *differential replication of memes*. Susan Blackmore illustrates the *memetic shift in perspective* that is connected to a memeticist understanding of culture: "Instead of thinking of our ideas as our own creation, and as working for us, we have to think of them as autonomous selfish memes, working only to get themselves copied"(Blackmore 1999: 8).[6] Aaron Lynch (whose book's merits lie not so much in theoretical analysis, but in its unique richness as a collection of memetic observations) illustrates his

[6] In his book on *The New Science of the Meme*, Richard Brodie tries to catch this "profound insight" in a sentence that reveals that, in a sense, it is not all that profound after all: "This is the most

slightly more moderate version of this "paradigm shift" (Lynch 1996: 17) with an example: "If a denomination expands, the sociologist usually asks what sort of advantages attract all the newcomers. The memeticist, on the other hand, studies the denomination's creed with an eye toward how it evolves and furthers its own replication" (Lynch 1996: 22).[7] Using yet another example, Daniel Dennett puts it more bluntly: "a scholar is just a library's way of making another library". Dennett continues, addressing his reader directly:

> I don't know about you, but I'm not initially attracted by the idea of my brain as a sort of dung heap in which the larvae of other people's ideas renew themselves before sending out copies of themselves in an informational Diaspora. It does seem to rob my mind of its importance as both author and critic. (Dennett 1993: 202)[8]

In the memetic view, *our* role is subordinate to the memes in a twofold sense: on the one hand, our mind is seen as the "dung heap" or "meme nest" (Dennett 1995: 355), i.e. the selective environment, in which the drama of the 'struggle for existence' of the memes takes place. At the same time, we are the meme's "survival machines" (Dennett 1995: 347–348), i.e. our mind is shaped or modified by the memes to provide for their survival.[9] Memes determine how we behave; thinking of Beethoven's Fifth Symphony might make us whistle and spread the tune to those around us. Or, to mention an example that has helped a great deal in popularizing the memetic account, the belief that the end is near is likely to "cause" proselytizing behaviour, which in turn favours further replication of that meme (cf. Lynch (1996).

In spite of the fact that, in many respects, memetics is a highly innovative and original project, and in spite of the fact that the program has attracted some rather prominent followers: thirty years after Dawkins' book, it might seem that the whole project was a failure. True, there are some sprouting twigs, especially in popular science – an example is Dean Hamer's book on *The God Gene* (2004), in which the author combines a genetic interpretation of spirituality with a memetic interpretation of organized religion. In spite of such examples, the general view has it that the attempt at infecting the scientific community with the idea of memetics has failed. On their homepage, the editors of the *Journal of Memetics* announce that they aim at re-launching the project; but this cannot detract from the fact that the last issue of the original journal dates from 2005, and contains a series of obituaries on the program. Susan Blackmore seems to have left academia completely, and is touring the world as an *infotainer*, spreading the memetic word to the general

---

surprising and most profound insight from the science of memetics: your thoughts are not always your own original idea" (Brodie 1996: 14).

[7] Lynch is unique among the memeticist in propagating a "non imperialistic" relation between memetics and other approaches, claiming that a kind of division of labour between different approaches is required (Lynch 1996: ix).

[8] Cf. also the almost identical formulations in Dennett (1995: 346). For a diagnosis of the "fundamental incompatibility" between the memetic approach and "the Cartesian voluntarism implicit in much social sciences", see also Marsden (1998).

[9] In Dennett's words, memes transform the "operating system" or "the computational architecture of the brain" (Dennett 1995: 340).

audience. Richard Brodie seems to have turned his back on memetics and become a professional poker player. After an episode of paranoia, Aaron Lynch died from an overdose of pain killers.

There is still some interest in memetics, especially on the part of philosophy and theory of science, with occasional papers on the topic appearing, and some conferences and collected volumes, but the general opinion seems to be clear. Considering the negative statements by the leading figures in the field, it becomes apparent just how much damage the memetic *shift of perspective*, i.e. the turn to the *meme's-eye view*, has done to the reputation of the entire program. Regularly, the critic's view is focused on that shift. To quote just one example: at the end of his review of a very academic and rich collection of papers on the topic that was edited by Robert Augner under the title *Darwinizing Culture* in 2000, John Dupré (2003) quotes a lengthy passage from Susan Blackmore's contribution, which contains the following statement of the memetic *shift of perspective:* "my inner self, which seems to have consciousness and free will, is in fact a memeplex created by and for the replication of memes. 'My' beliefs and opinions are survival tricks used by memes for their own perpetuation", only to then add the following disparaging remarks: "It is no doubt politest to pass over such nonsense in silence"; memetics, Dupré continues, is "theory going on holiday", a "simplistic idea" bloated to "quasi-philosophical nonsense".

As we have seen above, the memetic 'shift of perspective' from 'us' to 'them' (and the concept of the self as 'meme nest') is by no means a marginal notion within the memetic project. But still, it seems wrong to identify the whole project with this particular idea. Memetics without the *meme's-eye perspective* is possible. Before turning to this, however, another question has to be answered: what precisely is it that is wrong about the *meme's-eye* account of memetics?

In the following, I will first answer this question. In § 37, I shall argue that the memetic view rests on a distorted conception of the units of cultural evolution. The main problem of the memetic approach is that it ignores some fundamental ontological differences between DNA and 'memes'. It is built too closely on the model of genetic evolution, a model that is unfit for cultural evolution. In §38, Gabriel Tarde's idea of 'evolution by association' shall be introduced as a convincing alternative to the memetic idea of cultural evolution. In the concluding section (§39), I shall come back to the initial question concerning the place of the self in society: what is our role in cultural evolution in a Tardean view? Here, I shall take issue with some recent interpretations of Tarde's view of the self.

§37 Meme Ontology

As seen above, Dawkins' *gene's-eye perspective* narrative of biological evolution hinges on his peculiar notion of the gene. Dawkins does not endorse a functionalist conception of the gene, according to which genes are identified through their function within the organism. As this conception presupposes the organism as the whole

within which the gene is a (functional) part, this is at odds with Dawkins' notion of the ontological primacy of the gene over the organism. Thus Dawkins identifies the genes with the molecular structure of DNA. There is much to say in favour of a functionalist conception. But let's accept Dawkins' fundamentalism about DNA and ask the question: if genes are simply sequences of DNA: what, then, are memes? What *is* the replicator, the supposed new hero on the scene of the cultural world, really?

From an early stage in the development of the memetic movement on, the feeling that there was no clear answer to this question was pervasive among the participants in the debate (cf. Blackmore 1996: 92). In the relevant literature, memes are often defined by their method of replication: memes are units of imitation. This answer, however, simply begs the question. *What is it* that is imitated when somebody tells a story that she has read in the paper, or wears clothes according to the latest fashion, or unconsciously hums some tune which she has picked up in the elevator to her office, or uses the new recipe for cooking tomato soup that he has gotten from his mother-in-law?

In the received literature, there are two types of answers to this question: we might label them the externalist and the internalist answer, respectively. The externalist view defines the meme as a physical fact (cf. Gatherer 1998). In the case of the story and the melody, the meme is the sequence of sound waves, in the case of clothing, the meme is the texture of the fabric etc., in the case of the soup, the meme is the chemical composition of the liquid in question. This externalist view has the advantage of locating the meme firmly in the observable world. But it quickly runs into serious difficulties. It seems plausible that, in order to replicate, memes have to be physically manifest, be it only in the transitory form of sound waves or patterns of light. But a simple consideration reveals the limitations of this conception. Even though all memes have to be observable at a certain stage of their replication, the external physical manifestation of the meme has to be distinguished from the meme itself. This is apparent from the fact that the identity of the meme does not depend on any part of its external physical manifestation. The infamous story of the spider in the Yucca tree completely changes its external physical manifestation on it's way through the population. As printed in the newspaper, it's inkstains on cellulose. As told by people, it's sound-waves, which vary greatly from language to language. In spite of these differences, the story – the meme – is one and the same. The question is: what is it that remains the same in spite of the total difference in external physical manifestation?

The externalist view has no convincing answer to this question. Its internalist counterpart avoids this difficulty by defining the meme as something like the "unit of information in the brain" (Dawkins 1976). Thus the meme is not directly observable. But there are even more difficulties. What's the message of a melody? What kind of information do such memes convey? And to what extent is information "in the brain"? Is the meme ultimately some neurophysiological fact in the brain, i.e. not directly observable, but physical nevertheless? This seems implausible for the following reason. If a professor of English quotes Keats' *Ode on a Grecian Urn*, the neurophysiological state of his brain will very likely be very different indeed

from that of his students' brain citing the same poem. Still, the meme is the same. Thus the identity of the meme does not seem to be a matter of what goes on in the individual brain.

Neither the externalist view nor its internalist counterpart seems to be particularly satisfying, the reason being that both try to give the meme an ontological footing in the physical world (either outside or inside the brain). The alternative is to drop this preconception, and allow for a *interpretive* conception of the meme. According to this third, *interpretive* view (which I will advocate in the following), the moment of identity in the change of the physical "bearer" is the *meaning* of what is said, hinted at, or written. To see the meme in the physical or neurophysiological facts, it is necessary to take the "intentional stance" (Dennett 1995: 356). The meme is the *intentional object*.

Is the interpretive view adequate for all kinds of apparent memes? What about melodies: what do people *mean* by whistling a sequence of Beetoven's Fifth? It is not easy, and perhaps impossible, to distinguish between the sequence of sounds and some "meaning" in such cases. From this, it becomes evident that the interpretive view of meme is limited to a certain class of memes. We may call these memes the *symbolic* memes. But not all memes are of the symbolic kind. I conjecture that the term "meme" is equivocal, and that not everything that is discussed under this label can be adequately captured by one and the same theory. A systematic meme ontology would have to begin by sorting out these distinctions. This, however, is not my concern here. As it seems evident that symbolic memes play an important role, I will simply limit the following considerations to this class.

If the above considerations concerning an interpretive account of symbolic memes are correct, a fundamental critique of the memetic 'shift of perspective', and of its orientation towards the neo-Darwinian, DNA-foundationalist model of genetic evolution arises. The problem is this. Meaning is ontologically different from brute natural facts such as DNA-sequences: the ink lines on the paper and the sequences of sound waves have a *meaning* not in themselves, but only in relation to *somebody*. In this sense, the memes are *ontologically subjective*.[10] To use the examples John R. Searle gives to illustrate the difference between the ontologically objective and the ontologically subjective, memes (in terms of the 'meaning' of signs of any kind) are more like pains and aches than mountains or molecules. They are not there independently of whether or not there is somebody who is aware of them, which makes them different from things such as mountains that existed long before there was any consciousness of their existence. As social facts, symbols have a radically *subjective mode of existence*. It is only 'for us' ('us' in the sense of the members of the widest possible, and least parochial community of interpreters) that the scribblings and sounds that we use to communicate have the meaning they have. Thus it seems quite obvious that whoever talks about symbols, or institutional facts in general, cannot remain silent about 'us', the form of life whose conventions are constitutive of the institutional facts taken as the focus of memetic analysis. Memetic analysis has to take an 'intentional stance' towards us, because

[10] For an analysis of the basic traits of ontological subjectivity see Searle (1995: 7ff.).

memetics is ultimately about what we *mean* when we do things such as producing certain sounds, or making marks on the paper.

In his analysis of *Darwin's Dangerous Idea*, Dennett asks the decisive question concerning the ontology of memes: "what stands to a meme as DNA stands to a gene?" (Dennett 1995: 353). It seems that if, as Dennett himself seems to suggests, what 'makes up' the meme is ontologically subjective, then an unbridgeable rift opens up between memetic and genetic evolution, which has far-reaching consequences for the status of the memetic approach to culture. 'Our' relation to our genes is fundamentally different from the relation to our memes. In the case of memes, there is no equivalent to the DNA-sequences that floated freely in the primordial soup long before they began to build themselves 'survival machines' in order to survive in the struggle for existence. In contrast to DNA, meaning is ontologically subjective, and so our relation to memes is not that of a contingent 'meme machine'; beyond being the 'survival machines' the memes have built themselves, and beyond being the 'heap of dung' which more or less by accident is the 'ecological niche' in which memes thrive, we play much more important a role on the stage of cultural evolution. We are not simply 'meme machines' in the sense in which we might rightly be called 'gene machines'. And we are not just a part of the copying environment in which memes compete. Insofar as the relation to 'us' is an essential part of what meaning *is*, we are *constitutive* of the very 'matter' of which memes consist.[11] There is simply no meaning (and no memes) in the world without conscious beings having intentions, thoughts, and feelings, and without them being capable of interpretation, and mutual imitation, and of communication by means of the use of symbols. Thus it seems that, given the ontological difference between genes and memes, *serious doubts arise concerning the transfer of the genetic 'paradigm shift' to the memetic level*. As shown in section 36, this transfer is central to the memetic program. With the conclusions of the present section, this transfer seems to rest on a mistake. There is no equivalent to the neo-Darwinian, genetic shift of perspective from 'us' to 'them' in the cultural sphere. Independently of 'us' (i.e. the members of the widest community of communication), the memes would simply not exist. There is no equivalent to the primordial DNA (that is ontologically objective) in memetic evolution. Because it is ultimately about *meaning*, memetics is always and

[11] As is self-evident from what was said above, my use of the terms 'ontological subjectivity' and 'constitution' does not mean that I am committed to the view that meaning is 'produced' by some sovereign decisions of self-transparent subjective wills. In this context, 'ontological subjectivity' simply means that, for such 'brute facts' as sound waves, ink lines, and the like to have meaning, there has to be somebody around in relation to which those brute facts mean (or used to mean) what they mean (even though this 'somebody' might be utterly intransparent to herself and completely un-sovereign. More than self-intransparent, the subject in question might even be dead. For the latter case, take the example of 'indecipherable' signs from ancient ages, which still do have a meaning, even though there is nobody around who can tell us what that meaning *is*). Also, by calling meaning ontologically subjective, I am by no means committed to some radical idealist view according to which all events can be traced back to self-transparent subjects. Indeed I claim that such a view is a mere caricature of the 'traditional view', a mere fiction of some versions of the critique of the 'autonomous self' which were put forth about a quarter of a century ago, and which live a strange kind of after-life in some strands of current 'continental' social theory.

inevitably about 'us', i.e. the forms of life that make it possible to bestow things such as certain sequences of sounds with *meaning*. Memetics cannot abstract from those creatures that *make sense* of physical facts. This fundamental difference between 'our' relation to the (ontologically objective) genes on the one hand and 'our' relation to the (ontologically subjective) memes on the other drops out of sight because the memeticists are driven by the urge keep their theory of cultural evolution as close as possible to the neo-Darwinian model of genetic evolution.

As mentioned above, it was because he wanted the word to sound a little bit more like 'gene' that Richard Dawkins cut a syllable off the Greek word 'mimema' when he coined the term 'meme'. This episode concerning its origin (Dawkins reports it himself) epitomizes what is problematic about the whole memetic venture, wrapping up nicely the fatal tendency of memeticists to model cultural evolution on genetic evolution. Just as Dawkins simply cuts off that part of the word that did not seem to fit, memeticists distort the phenomenon to fit the model of genetic evolution. Once one becomes aware of this problem, the question remains to be answered: what will the memetic research program have to turn into, once the orientation towards the model of genetic evolution (which is so fundamental and so fatal for the memetic account at the same time) is given up? What does an adequate understanding of cultural evolution look like?

§38 Evolution by Association

The importance of Tarde's *Laws of Imitation* (Tarde [1890] 1921) for the issues at stake in the controversies around the memetic project has not escaped the notice both of memeticists and of their critics. Even though there is no systematic analysis of the topic available as yet, it seems that there are three typical views of the relation between memetics and Tarde's theory of imitation. The first view is expressed in a paper that has stirred much of the current 'Tardomania'. In his paper 'Gabriel Tarde and the End of the Social', Bruno Latour claims that the memetic account (Latour refers to Blackmore) is just a 'simplified version' of Tarde's monadology, thereby implying that, on the fundamental level at least, there is no real disagreement between these theories (Latour 2002: 119–120). The second view concurs with the first, but gives a different twist to the diagnosis of some 'fundamental agreement'. Looking from the 'memetic' side of the relation in question, Paul Marsden has made a similar remark concerning some deep affinity – even though in his 'memetic' view, the integration between Tarde's theory of imitation and Memetics goes the other way around. Following Marsden, Tarde's "programme for sociology" has so "much in common with the memetic project" that, in spite of his deplorable failure to grasp and appreciate the essential features of Darwinism and in spite of his somewhat less concise concept of imitation, Tarde should be honoured as one of the most important "Forefathers of Memetics" (Marsden 2000). The third view (which is closest to the one to be developed in the following) does not deny the similarities either. Here, Tarde is credited with having anticipated almost everything that is

interesting about memetics. At the same time, however, a fundamental difference between Tarde and memetics is emphasized. In his critique of the memetic project, Gustav Jahoda focuses on the memetic 'shift of paradigm' or 'shift of perspective' which, in his view, seems to be the main mistake of memetics, especially of Blackmore's interpretation. "Intentionality has been transferred away from humans and to the memes" (Jahoda 2002: 65). According to Jahoda, his reluctance to make this move is what makes an alleged "Forefather of Memetics" like Gabriel Tarde superior to the memeticists themselves.

Before addressing this particular issue, let's take a closer look on Tarde's relation to Darwinism. Tarde's paper on 'Darwinisme naturel et darwinisme social' from 1884 provides a vivid illustration of some of the problems that, at first glance at least, faced any adaptation of pre-genetic Darwinian thinking for the purpose of the social sciences. Paradoxically, the most basic problem does not arise from the biological orientation of Darwin's theory. Rather, it stems precisely from those motifs, which Darwin had taken over from social science. In Tarde's view at least, Darwin's fatal mistake was to follow the Manchester School of Economics in its obsession with (and fixation on) the "magic power of competition" ("vertu magique de la concurrence", Tarde 1884: 614). Tarde is well aware of the fact that the Darwinian motifs of the 'struggle for existence' and the 'survival of the fittest'[12] originated in the theory of capitalism. And it is on these grounds that Tarde criticizes Darwinian 'selectionisme'. Following Tarde, the problem of Darwinism is not the transfer from the 'social' to the 'natural' or vice versa, as one might think from the title of Tarde's paper, but the fact that Darwin has made the wrong choice in modelling the theory of evolution on a distorted theory of economic competition. In Tarde's view, competition, rivalry and conflict are only one of the two fundamental types of social relations, the other being *cooperation*, which in economical terms corresponds to labour ("travail"; cf. Tarde 1884: 614). For Tarde, Darwin's "Manchester school-like" tendency to overestimate the role of rivalry and competition at the cost of the role of cooperation led him to neglect the very preconditions of any 'struggle for existence': "one has to be strong in order to fight, and strength comes from interior unity" (Tarde 1884: 613). In this view, any 'struggle for life' *presupposes cooperation*. What makes the individual body 'fit' is the cooperation between the organs. Thus, contrary to Darwin's account, *cooperation* rather than *competition* should be the prime topic of analysis. Tarde sums up his criticism of Darwinism, when he proposes to replace Darwinian 'selectionisme' by an understanding of 'évolution par association' (Tarde 1884: 613).

Many of the insights that put Tarde back on the map of current social theory (after a rather long 'latency period'), are directly connected to this theory of 'évolution par association', especially the idea that every individual is a society, and the thesis that any whole is always less complex than its parts (in his paper, Tarde presents both of these views as direct objections to Darwin; Tarde 1884: 609). However, I will not pursue this particular aspect of Tarde's 'évolution par association' any further here. Independently of the question of whether or not this criticism, as directed against

[12] For these two core concepts of the Darwinian account, see Darwin ([1859] 1975), chaps. 5–6.

Darwin, was justified (at first sight, at least, it seems that it was), it clearly seems that genetic neo-Darwinism does not adhere to the old metaphysics of the development from simple elements to complex wholes anymore, thus making a renewal of the Tardean line of critique somewhat redundant. Neo-Darwinians are generally aware of the communal character of wholes, and, connected with this, of the complexity of lower levels of integration. As Richard Dawkins puts it, any individual has basically a "communal character" (Dawkins 1976: 25): it is the (more or less) cooperative project of a large multitude of genes. And along these lines, nothing seems to speak against seeing the genes themselves, in turn, as cooperative projects to provide for the preservation of the even more complex molecular protein structures that make up the genes. Thus it seems that modern, genetic evolutionary theory has long integrated the basic insight of Tarde's 'évolution par association', the insight that any higher-level competition (e.g. between individuals) goes hand in hand with lower level-cooperation (e.g. between genes). It seems that what Tarde harshly criticizes under the title of *selectionisme* has not much to do with Neo-Darwinian thinking in the first place. Indeed it might even appear that Tarde's 'evolution by association' fully conforms to the neo-Darwinian memetic program. For Tarde himself identifies 'association' (or 'travail') with imitation (Tarde 1884: 615), and famously, Tarde compares the role of imitation for the social with that of inheritance in the biological sphere (both are forms of 'universal repetition', the third being 'ondulation' within the physical sphere; see Tarde [1890] 1921: 1ff.). This closely parallels the memetic idea that imitation and inheritance are the two 'ways of replication' in the memetic and genetic evolution, respectively. Beyond the idea of replication, Tarde's theory of imitation also seems to include *variation* and *selection*, the other two essential features of any theory of evolution. According to Tarde, variation is an essential feature of any imitation: "actually, even the most imitative of all men is innovative in some respect" ("à vrai dire, le plus imitateur des hommes est novateur par quelque côté," Tarde [1890] 1921: ix). And when such innovations in turn become the object of imitation, they are not only subject to all sorts of *recombinations* and mutual *reinforcements*, but sometimes stand in a direct 'struggle for existence' against each other. Tarde captures this in a concept he takes over from the physics of waves, when he speaks of the various phenomena of 'interference' between imitations, and he gives a detailed description of what parallels the role of selection in the theory of evolution under the heading *le duel logique* (Tarde [1890] 1921: 167–187). At the same time, the relation between Tarde's theory of imitation and the memetic account of cultural evolution is not exhausted in close analogies that are somewhat obscured by a few minor misunderstandings of the essence of Darwinism on Tarde's side. Tarde, who is not obsessed with the memetic idea of modelling cultural evolution on genetic evolution, gives a surprisingly clear and convincing answer to the question concerning the nature of the content of imitation:

> What is imitated is always an idea or a wish, a judgement or a plan, in which a certain amount of *belief* and *desire* are expressed, which is the entire soul of the words of a language, the prayers of a religion, the administrations of a government, the paragraphs of a

> code of law, the duties of a moral system, the work of an industry, the products of an art. (Tarde [1890] 1921: 157)[13]

The ultimate 'objects of imitation' are our beliefs and desires. This requires a somewhat more ambitious concept of imitation than the one that is commonly used. In Tarde's words, imitation does not go "outside in", but "inside out". It does not pick up what is observable about the actions of others first. Rather, the imitation of other people's expressions and behaviour comes only *after* the imitation of their ideas.[14] This understanding of imitation goes much against the grain of Stephen Turner's recent attempt to turn Tarde into the leading figure of a whole alternative paradigm in social theory, because the latter is based on an 'externalist' understanding of imitation, in which it is only and exclusively the 'outside aspects' of actions which are imitated.[15] At the same time, Tarde's 'internalist' theory of imitation seems to offer a convincing solution to the above-mentioned problem that faces the memeticist account. *Tarde's approach to cultural evolution does justice to the ontologically subjective character of the 'replicators'*. If 'evolution by association' is about our beliefs and desires, it is, as a matter of course, always and inevitably about *us*, i.e. about the kind of creatures whose beliefs and desires cultural evolution is all about. Thus Tarde's theory of 'evolution by association' avoids the fatal memeticist tendency to displace the 'self' by the meme. This does not mean, however, that Tarde returns to the subject as 'author' and sovereign in the cultural sphere. The Tardean view on the role of imitation does not provide support to the anti-memetic thesis that the assumption of a 'meme's-eye perspective' (in terms of an inherent tendency of the memes to replicate) is unnecessary because memes are copied only insofar as they seem to be useful for the projects of *persons* (and not because replication

---

[13] *"Ce* qui est imité, c'est toujours une idée ou un vouloir, un jugement ou un dessein, où s'exprime une certaine dose de *croyance* et de *désir*, qui est en effet toute l'âme des mots d'une langue, des prières d'une religion, des administrations d'un État, des articles d'un code, des devoirs d'une morale, des travaux d'une industrie, des procédés d'un art" (Translations from *Les lois de l'imitation,* as well as from 'Darwinisme naturel et darwinisme social', are mine). Shortly after the above-quoted passage, Tarde argues that, beyond the apparent duality of belief and desire, the ultimate object of imitation (and thus the essence of society) is *belief,* because, in the last resort, belief (in the form of convictions) is what desire is all about (cf. Tarde [1890] 1921: 160, where Tarde provides some further explanations). This view seems to reflect the classical 'Cartesian' primacy of cognitive intentionality (cognitive attitudes such as beliefs) over practical intentionality (practical attitudes such as intentions), as criticized by Martin Heidegger in *Being and Time* ([1927] 1996: §12–13). Without being able to argue at sufficient length here, I see no real reason to follow Tarde on this reductive move.

[14] Cf. Tarde ([1890] 1921: 225): "Cette marche du *dedans* au *dehors*, si l'on cherche à l'exprimer avec plus de précision, signifie deux choses: 1° que l'imitation des idées précède celle de leur expression; 2° que l'imitation des buts précède celle des moyens. Les *dedans* sont des buts ou des idées; les dehors, des moyens ou des expressions."

[15] Cf. Turner (2000: 106): "Imitation is wholly external: one can imitate only what one can see or hear, that is to say, the externals of an act, thus the content of imitation, are limited by our ability to identify something to copy. We may imitate unconsciously, but this does not mean that we have special powers of unconscious discernment that allows us to discern anything other than the external aspects of what we imitate."

is good for the meme's own project).[16] In a Tardean view, imitation is not about copying what seems useful for one's projects. Rather it is, *the projects themselves* that are copied (and with it the standards by which usefulness is measured).

§39 Hypnosis Versus 'Openness to the External World'

Thus the role and concept of the self in the Tardean theory of socio-cultural evolution is different both from the memetic view and from the concept of the self as the sovereign 'source' and 'author' of the cultural. From a Tardean perspective, the self is neither the selective environment ('heap of dung') nor the 'survival machine' of memes, let alone the 'wrong idea of the self' against which the memeticist critique is levelled.

But what is it, then? In a famous passage of his *Laws of Imitation*, Tarde compares the state of the subject in society to that of a *hypnotized* individual.

> The social state, just as the hypnotic state, is a dream of control and a dream in action. It is the illusion of both the somnambulist and the social human being alike to take those ideas to be spontaneous, which in fact she has taken over by suggestion. (Tarde [1890] 1921: 83)[17]

In this sense, the fact of imitation (which, according to Tarde, is the essence of the social) seems to run counter to our alleged 'intentional autonomy'.[18] If beliefs and desires are ontologically subjective in the sense that there has to be someone who 'has' the beliefs and desires in question, it is also true that these beliefs and desires are not really *hers* or *his*. Famously, Tarde ([1890] 1921: 266ff.) criticizes the allegedly 'enlightened' self-understanding of modern men who think they have freed themselves from the old authoritarian structures, and are following only their own best judgement. In Tarde's view, this vision of the sovereign enlightened self is purely illusory, for all that has changed between the middle ages and modernity is that people now mutually hypnotize each other, instead of being hypnotized by some leading figures or traditions. Thus the Tardean version of cultural evolution, just like the memetic version, requires some 'shift of perspective'. However, in the case of the Tardean model of cultural evolution, this 'shift of perspective' is no shift away from 'our intentionality' to the 'meme's-eye view', as it is in the memetic project. Tardean evolution is not just about 'memes' in terms of thoughts and desires. The Tardean version of the evolutionary 'shift of perspective' is much more complex than the memetic one. Instead of changing the cast of the evolutionary play, instead of replacing our intentionality as the hero at the centre of the stage

[16] For such a view cf. Millikan 2003: 105–106).

[17] "L'état social, comme l'état hypnotique, n'est qu'une forme du rêve de commande et un rêve en action. N'avoir que des idées suggérées et les croire spontanées: telle est l'illusion propre au somnambule, et aussi bien à l'homme social."

[18] For a powerful holistic defence of intentional autonomy against the collectivist alternative that has been of tremendous influence in analytical social theory see Pettit (1996).

with some new agent (i.e. the meme), the scene is left intact, but it is shown in a wholly new and different light. Our intentions (beliefs, desires and affective intentions) are not viewed in the perspective of their "direction of fit" (to use Searle's terminology; cf. Searle 1983). It is not (or not primarily) the *truth* of our beliefs or the *fulfilment* of our desires, i.e. the relation between the intentional subject – the one who 'has' the intention – and the intentional content, i.e. whatever the intention in question is *about*, which is highlighted in the Tardean perspective. The social side of our intentions is not to be found in the relation between the intentional subject and the intentional content, but in the relation between different intentional subjects. What makes beliefs and desires the object of analysis in social science is not the fact that these intentions do (or do not) refer (or meet their conditions of satisfaction), it is not a matter of what these beliefs are *about* (and of whether the beliefs in question are true or not). Rather, it is a matter of how these beliefs spread (or vanish) in a population. Thus the perspective of the theory of 'evolution by association' cuts across our everyday perspective. Whereas normally our intentions are relevant to us in the perspective of what they are *about*, the Tardean view looks at from where (or whom) these intentions are *taken over*, and to whom (and how) they are *passed on*.

For a more precise understanding of the relation between these two divergent perspectives (which might be called the 'referential' view and the 'social' view, respectively), it is important to keep in mind the following. Tarde criticizes the view of our individual self as the source of all of our intentionality, a view which is proven wrong by the fact that most of our beliefs and desires are by no means our own individual creations, but in fact taken over from others. However, Tarde does not claim that the question of whether or not intentions meet their conditions of satisfaction (i.e. whether or not beliefs are true) is *irrelevant* for an understanding of our intentionality (i.e. our beliefs and desires). And the fact that imitation is the way by which most beliefs are acquired does not mean that the beliefs and desires that are the object of Tardean analysis do *not* refer to an external world. True, this is what Tarde seems to suggest when, in the most famous passage of *Laws of Imitation*, he characterizes the "social man" as a "somnambule" (Tarde [1890] 1921: 83). It is tempting to interpret this passage as saying that it is not our relation to the external world, but our relation to our 'co-believers' that determines what we believe. Upon a closer reading, however, it becomes apparent that Tarde's view between the 'social' or 'imitative' character of our beliefs, on the one hand, and our intentional 'openness' to the external world (an expression which Tarde occasionally uses himself!)[19], on the other, is much more complex than that. In order not to misinterpret Tarde's somnambule, it is crucial to read closely:

> Assume a man who, *by way of hypothesis, is stripped of all extra-social influence, of the direct sight of natural objects, of the spontaneous obsessions of his different senses, and has no communication but with his fellow human beings* . . . Is not this the object suited for the

[19] Cf. e.g. Tarde ([1890] 1921: 86), where he speaks of the individual human being as a "natural being, susceptible and open to the impressions of the external nature" ("être naturel, sensible et ouvert aux impressions de la nature extérieure"), thus contrasting the 'natural' intentional openness with the 'social' somnambulism.

> study, through experience and observation, of the essential features of the social relation, thus *detached of all influences of natural order and physics* . . .? (Tarde [1890] 1921: 83)[20]

In other words, Tarde's somnambule is an *abstract* entity, i.e. the result of an abstract thought experiment. The fact that our beliefs and desires are taken over from others reveals us as *social* beings. This, however, does not mean that the reference to the world ("direct sight of natural objects") and the spontaneity of some elementary form of desires ("spontaneous obsessions") are inexistent in (or inessential to) our intentionality. Tarde does not mean to claim that it is only as *social* beings that we have beliefs and desires. He just states that, very often, our intentionality has a *social* aspect, which consists in the fact that our beliefs and desires are taken over from others, and that if and insofar as this is the case, the intentionality in question cannot be analyzed *exclusively* in terms of the relation between the 'subject', on the one hand, and the 'object' which the intention is 'about', on the other. In this case, the relation to other 'intenders' (which runs across the intentional subject–object-relation) has to be taken into account. And this is what social science is supposed to do.

Thus it seems that even though Tarde's image of the *social* man (the somnambulist) seems to fly into the face of the modern human self-image, his *overall view* of our beliefs and desires is compatible with a great deal of *intentional autonomy* in terms of spontaneity of desires and intentional openness to the external world. (As we shall see below, Tarde even suggests that our intentional openness is a *precondition* of our social somnambulism.)

Before taking a closer look at the relation between intentional openness and somnambulism, a critical remark concerning the current 'Tardomania' (Mucchielli 2000) is in order. The fact that Tarde characterizes his somnambulist, i.e. the purely imitative dimension of our beliefs and desires, as the result of an *abstract* thought experiment puts him directly at odds with those of his present-day interpreters who like to see the imitative dimension as the 'foundation' of our beliefs and desires (typical for this view is Leys 1993). As the above-quoted passage indicates, this is not what Tarde suggests. Tarde introduces his somnambulist quite explicitly as an abstract entity. Abstractions, however, are generally rather unfit for foundational purposes of any kind. For any abstraction *presupposes* whatever it is that is left out of the picture in the course of the abstractive process. In this sense, Tarde's picture of the *somnambule* as the 'social self' *presupposes* the sense in which social *and* 'extra-social' influences are closely intertwined (indeed, as we shall see below, Tarde claims that the social self has its roots in the pre-social sphere). This, however, is at odds with some recent attempts to pit Tarde against one or another of the more 'traditional' views of the role of the self. Contrary to what Ruth Leys seems to think, Tarde does not "break with the Cartesian ontology of the

---

[20] "Supposez un homme qui, *soustrait par hypothèse* à toute influence extra-sociale, à la vue directe des objets naturels, aux obsessions spontanées de ses divers sens, n'ait de communication qu'avec ses semblables . . . [N]'est-ce pas sur ce sujet de choix qu'il conviendra d'étudier, par l'expérience et l'observation, les caractères vraiment essentiels du rapport social, *dégagé ainsi de toute influence d'ordre naturel et physique . . .?*" (my emphasis).

autonomous subject by defining the self in terms of the social" (Leys 1993: 282). As we have seen, Tarde is well aware of the extra-social dimensions of selfhood even when he draws our attention to the social aspects. What he does is something quite different from breaking with the 'Cartesian ontology of the autonomous subject': he adds a *social ontology* to the Cartesian ontology of the mind. At first glance at least, it might even seem that Tarde is not completely free of the under-socialized (or perhaps even anti-social) Cartesian image of intentionality, according to which anything that is 'valid', 'true' and 'authentic' about our intentionality is basically a matter of the monological representation of the external world in the secluded immanence of the individual mind. In the Cartesian view, the self is 'social' only insofar as it is disturbed and deflected by 'social' factors such as authority, traditions and conventions. Tarde is much closer to this view than one would guess from his current reception. Thus he explicitly states that it is only because we are not just 'social' beings, but also 'natural' beings, that we can *change* and *renew* our lives and culture: "If the social being was not at the same time a natural being, sensitive to the impressions of external nature [. . .] he would not be susceptible to change"(Tarde [1890] 1921: 86).[21]

More than this, in a passage that faintly echoes Descartes's departure from social life in his "lonely withdrawal"[22] to his secluded castle in his *meditationes*, Tarde says that to make inventions or discoveries, we have to break away from our groups and societies: "In order to innovate, in order to make discoveries, in order to wake up for a moment from her or his familial or national dream, the individual has to escape from her or his society for a moment" (Tarde [1890] 1921: 95).[23] Upon closer examination however, it becomes apparent that the relation between the 'natural' and the 'social' aspects of our intentionality is not as static as it may first appear. Tarde makes clear that there is no clear-cut line between the 'natural' and 'social' aspects of beliefs and desires. It is not the case that our beliefs and desires are *either* to be located in our 'sensual' relation to the world *or* taken over from others by means of imitation. Rather, our natural 'openness to the external world' (as "natural beings, susceptible and open to the impressions of external nature"; Tarde [1890] 1921: 86), on the one hand, and our 'hypnotized' state as social 'somnambules', on the other, are in a close interplay. By way of imitation, beliefs and desires are 'refined' both by new inventions and by other imitations (Tarde [1890] 1921: 159). In this sense, new discoveries about 'external nature' are often made on the base of beliefs, which are taken over by way of imitation. At the same time, Tarde makes it clear that the 'natural' is more basic than the 'social' because our 'openness to the external world' logically *precedes* our 'somnambulism'. All beliefs and desires are ultimately rooted in the pre-social dimension of our intentionality: "These beliefs and desires, which are specified (and in this sense created) by invention and imitation, *have their deeper*

[21] "Si l'être social n'était pas en même temps un être naturel, sensible et ouvert aux impressions de la nature extérieure . . . il ne serait point susceptible de changement."

[22] Cf. Descartes' *Meditationes de prima philosophia*, 1st Meditation, §3.

[23] "Pour innover, pour découvrir, pour s'éveiller un instant de son rêve familial ou national, l'individu doit échapper momentanément à sa société."

source beneath the social world, in the living world" (Tarde [1890] 1921: 159; my emphasis).[24]

In this sense, the 'referential' aspect precedes the social aspect of our intentionality. The referential 'aboutness' of our intentionality, i.e. our intentional 'openness to the external world', is more fundamental than the imitative character of our beliefs and desires. Thus the picture of our 'somnambulism' has to be relativized. If we take over other's beliefs and desires by means of imitation, this is only *insofar* as and *because*, as 'natural' (i.e. pre-social) beings, our intentionality already entails a genuine 'openness to the external world'. This directly relates to the question of the role of the self in Tardean 'evolution by association'. The following overall view emerges. Our intentional 'somnambulism' is ultimately founded in our 'openness to the external world'; in most cases, the two 'aspects' of our intentionality (the 'referential' or 'natural' and the 'social') do not conflict with each other, but stand in a relation of mutual refinement and specification. At the same time, Tarde is well aware of the fact that there are exceptions to the rule. The 'hypnotic' power of certain beliefs to conquer our minds by way of imitation does not always go nicely hand in hand with the truth of these beliefs. Thus, in his paper *Darwinisme naturel et darwinisme social*, Tarde describes the Darwinian selectionist idea of the 'survival of the fittest' as a "magic formula which has the gift to capture the mind which it enters". However, Tarde does not think that the spread of this idea is in any sense inevitable. Indeed, he calls upon his readers to beware the "bewitching power" of this and similar ideas ("méfions-nous de leur ensorcellement" Tarde 1884: 607). Thus Tarde addresses us not *only* as the social 'somnambules', who believe whatever they are told, but *also* as 'natural' beings who, to some degree at least, are able to resist those 'contagious thoughts' by means of a critical assessment of their *truth or falsity*.

I conclude with a brief list of the main arguments of this chapter. (1) By contrast to the DNA-oriented view of the units of biological evolution (genes), the units of cultural evolution ('memes') are ontologically subjective. (2) This is ignored in the memetic view, which heavily distorts the theory of cultural evolution to fit the all-powerful, DNA-fundamentalist model of genetic evolution. (3) The mistaken analogy between gene and 'meme' leads to a mistaken view of the role of the self in cultural evolution. (4) Gabriel Tarde presents us with a theory of cultural evolution that is not flawed by the memetic orientation on genetics. (5) Tarde's concept of the 'social self' does justice to the social aspects of our intentionality, without being incompatible with a robust conception of the objectivity of our intentional states and of our intentional autonomy. This is overlooked in much of the recent "Tardomaniac" literature.

[24] "Ces croyances et ces besoins, que l'invention et l'imitation spécifient et qu'en ce sens elles créent . . . *ont leur source profonde au-dessous du monde social, dans le monde vivant*."

Chapter 12
Consensus

Learning from Max Weber's Problem

Max Weber's status in social science in general, and in German social theory in particular, is ambivalent. On the one hand, his importance as one of the foremost classics of social theory and social science is uncontested. On the other hand, however, he is routinely accused of relying on a skewed methodology. Especially, Weber's claims concerning the action theoretic foundations of sociology have been criticized. Because Weber's action theory is at the very heart of his work, and cannot be separated from his sociological theory, this is no insignificant charge.

In the development of German social theory, particularly during the last decades of the twentieth century, this diagnosis has played a crucial role. During this time, the field was divided into two camps, and the interpretation of Weber's work served as one of the battlegrounds for their controversies. The first camp's label was sociological Systems Theory, with the later Talcott Parsons and Niklas Luhmann as its main protagonists. The second camp was gathered around the label "Critical Theory", with Jürgen Habermas as its theoretician-in-chief. Both camps claimed to be able to provide a solution to the problem of Weber's action theory, while accusing the other camp of failing miserably at this task.

Much ink has been spilled on the controversy between the two camps, and the very peculiar role Weber seemed to have played in this has not gone unnoticed, either. Yet it has not been noticed thus far that there is one particular Weberian category that plays a crucial role in this. It is a category that, even in the enormous body of literature on, Weber does not seem to have received the attention it deserves: the category of *Einverständnis* and *Einverständnishandeln*, which Weber analyzes in the sixth section of his essay *On Some Categories of Interpretive Sociology* which was published in 1913.[1]

Einverständnis and *Einverständnishandeln* are no easy terms to translate. In the existing translations, the former is usually rendered either as "agreement" or as "consensus", while the term "consensual action" seems to be the preferred choice as far as *Einverständnishandeln* is concerned. As always, each of these translations has its problems. "Agreement" seems to have the advantage of being more

[1] Weber ([1913] 1981). German original: "Über einige Kategorien der vertsehenden Soziologie" (1913) in Weber (1922): *Gesammelte Aufsätze zur Wissenschaftslehre*, hereafter quoted as WL.

H.B. Schmid, *Plural Action: Essays in Philosophy and Social Science*, Contributions to Phenomenology 58, © Springer Science+Business Media B.V. 2009

common in ordinary English language than "consensus", but in order not to sever the conceptual link to "consensual action", I shall either use "consensus", or the German original in the following.[2]

The term *Einverständnis* plays a dual role in Weber's action theory. On the one hand, it stands for a basic phenomenon of social reality. On the other hand, it indicates a fundamental problem for Weber's action theory. To put it in Weberian terms (which will be explained below), the *phenomenon* is this: there is a special case of social action in which the action, in its "subjectively intended meaning", is oriented towards other agents' actions in the form of *normative expectations*. Now why should that phenomenon be a problem for Weber's theory? It is this: how can this phenomenon (and with it the role of social normativity in general) be accommodated within an action theoretical framework that assigns a paradigmatic role to instrumental rationality and goal-oriented action? Or, to put the problem more simply: why should rational instrumental agents ever *have* such expectations?

Parsons, Luhmann, and Habermas unanimously believe that the problem cannot be solved, because the phenomenon simply goes beyond the conceptual capacities of Weber's (or, in Habermas' diagnosis: the *official* Weber's) action theoretical framework. And both camps – Systems Theory as well as Habermas' version of Critical Theory – draw their consequences by departing from Weberian action theory and by basing their respective theoretical edifices on new foundations. In Parsons' and Luhmann's case, this is done by switching from an action theoretic framework to systems theory. Habermas, for his part, disavows Weber's instrumentalism and intentionalism, and turns to a linguistic account that assigns the paradigmatic role to communicative action rather than to goal-oriented action.

Both camps claim to have solved Weber's problem. Interestingly, however, they deny the other party any recognition of the achievement they claim for themselves. Thus, in Luhmann's view, Habermas simply remains stuck in the old action theoretic conceptual framework within which it is simply impossible to deal with the problem, whereas Habermas makes no secret of his view that Luhmann, with his functionalism, fails miserably to overcome the mentalistic instrumentalism of Weber's theoretical framework.

Given this constellation, and the importance it had for the development of social theory, it might be worth the effort to take a closer look at the source of all this trouble. What precisely is the phenomenon, and why exactly does it not seem to fit into Weber's theoretic framework? And, above all: how can Weber's problem, if there is any, be solved: is it the Habermasian or rather the Luhmannian solution that works – or neither of the two? And if so: is there a third way?

[2] Wherever possible, my translations will follow Richard Swedborg's *Max Weber Dictionary* (2005).

§40 The Problem of Interaction

In his essay *On Some Categories*, as in his other writings, Weber claims that the proper object of interpretive social science is a particular kind of action. In Weber's account, action is defined as intended behavior, i.e. behavior that has some "subjectively intended meaning". For an action to be the proper object of interpretive social science, it has to meet the following requirements:

> Action significant for interpretive sociology is (. . .) behavior that (1) in terms of the subjectively intended meaning of the actor, is related to the *behavior of others*, (2) is *codetermined* in its course through this relatedness, and thus (3) can be intelligibly *explained* in terms of this (subjectively) intended meaning (Weber 1981: 152).

Weber labels this kind of behavior "*Gemeinschaftshandeln*", a term which he uses elsewhere in his work, and which is usually translated as *communal action*. In the only existing translation of the essay *On Some Categories*, however, the term is rendered with "social action", which also translates Weber's *soziales Handeln*. What is the reason for this? First, it is obvious from Weber's definition that, in the essay *On Some Categories*, the term *Gemeinschaftshandeln* is used in the very same sense as the term *soziales Handeln* has in the rest of Weber's work, and especially in Weber's *magnum opus*, which was posthumously edited under the title *Economy and Society*. Weber also uses the term *Gemeinschaftshandeln* in *Economy and Society*, but he does so in quite a different meaning. *Gemeinschaftshandeln* is here defined as a small subclass of *soziales Handeln* (a type of action which is marked by the fact that the agents are tightly connected by a sense of belonging together).[3]

This latter meaning, i.e. the use of the term *Gemeinschaftshandeln* in *Economy and Society*, seems to correspond much better to the intuitive notion of "communal action." As we shall see, the meaning in which Weber uses the term *communal action* in his essay *On Some Categories* might appear highly counterintuitive. But still, we should not ignore that he does use the term here, and we should, I think, resist the temptation to correct Weber's terminology. Especially since the following interpretation is largely based on Weber's essay *On Some Categories*, I will use Weber's terms as he defines them in that essay. I will translate *Gemeinschaftshandeln* literally, i.e. as "communal action", but it is important to remember that this term is equivalent to "social action" as defined in *Economy and Society*. I will not talk about the narrow definition Weber gives the term "communal action" in *Economy and Society*, so I will only use that term in the sense of Weber's earlier essay.

Why does that term seem so strange that even the translator of the essay chooses to replace it with "social action"? The reason is simple. One of the most obvious characteristics of communal action, as defined in *On Some Categories*, is that it does *not* presuppose any community. Therefore, communal action simply seems to be a *misnomer*. As quoted above, communal action (in the sense of the essay

[3] In *Economy and Society, Gemeinschaftshandeln* is as a type of primary group action "based on the subjective feeling of the parties, whether affectual or traditional, that they belong together". Cf. Graber (1981).

On Some Categories) simply requires that some agent's behavior be oriented, in its subjectively intended meaning, towards another agent's behavior. It is true that it takes at least two for an action to be communal – there has to be at least one other agent around, to whose behavior the agent is oriented – but it seems that there does not have to be any community, or, in Weber's terminology, any form of *social relation* between the two whatsoever. Assume that A is the agent, and B is the other person towards whose behavior the "subjectively intended meaning" of A's behavior is oriented. For there to be communal action (or, in the terminology of *Economy and Society*, social action) as defined by Weber, it is not necessary that B in any sense *participate* in the matter. Indeed, B does not even have to *know* what the "subjectively intended meaning" of A's behavior is, or indeed that there is another agent. This is what makes talk of *communal action* particularly counterintuitive. It is true that the examples Weber uses are usually of a different and more communal character, but if we stick to the letter of the definitions Weber gives us, there need not be any reciprocity whatsoever for there to be communal action. B might think he is all alone on the planet; yet, as long as A, who is creeping up behind him to mug him, effectively orients the meaning of his behavior to B's, his action is "communal action", even though there is nothing communal about that action in the intuitive sense of the term.

As far as I know, there is no clear indication to be found anywhere in Weber's work that Weber ever thought this definition of the primary object of interpretive social science to be fundamentally deficient. But there are some hints that he thought that some specifications were in order. Thus, in the third paragraph of *Economy and Society*, Weber focuses on what he calls "social 'relation"'. In the case of social relations, the orientation towards the other agent's behavior is *reciprocal*. Social relation is defined as the "actions of several persons that are mutually adjusted and oriented to each other in their meaning (...). A minimal degree of relation of *reciprocal* orientation (...) is an essential trait of the concept". Most certainly, it is an advantage of Weber's concept of social relation, rather than any shortcoming, that it does not favor *cooperative* kinds of social relations over *competitive* kinds. "Enmity" and "competition" figure prominently on Weber's list of social relations, among many cooperative examples. Yet there is something else that should be noted, and that should be carefully distinguished from the fact that Weber leaves room for conflict-laden forms of social relations. Weber also allows for the possibility that, for all the mutuality and reciprocity of the subjectively intended meaning, social relations might nevertheless be "objectively 'one-sided"', as he puts it. This is the case where the agents, as Weber continues, "attach different meanings to their behavior." Weber's further explanations make the reader think that Weber, still under the label of "social relation", even wants to allow for cases in which there is mutual misunderstanding, i.e. in which the agents mutually ascribe attitudes to each other which they do not actually have. This is just another effect of Weber's basic action theoretic commitment, according to which the "subjectively intended meaning" that is constitutive of action need not be "true" or "correct" at all. The consequence for the concept of social relation is this: any trace of actual reciprocity is stripped off that concept; objective reciprocity is, as Weber says, simply a "limiting case" of social relation. Thus it seems even social relations need not in any sense be communal.

Weber thinks that social relations are simply a special case of the meaning structure of social action. Upon closer consideration, however, it seems that the step from social action to social relations is a decisive one for the entire project – at least as far as Weber's claim that "interpretive sociology (…) is not part of psychology" (Weber 1981: 154) is to be taken seriously. The step is from the analysis of single agents' actions to the "behavior of a plurality" (*Sichverhalten mehrerer*), as Weber says in §3 of *Economy and Society*. The question to be addressed in the following is this: can this transition really succeed within Weber's action theoretical approach? As it shall turn out, the category of *consensus* plays the key role in this transition.

But first, it has to be shown why this transition should be a problem for Weber at all. In order to do so from an *immanent* perspective, it is necessary to give a somewhat broader account of Weber's theory of the structure and role of subjectively intended meaning in paradigmatic cases of action. We have already introduced the terms "communal action" (or *social action* in the terminology of *Economy and Society*) and *social relation*. Let me briefly introduce the most important Weberian distinctions of action types (where not otherwise indicated, the following quotations are from the first paragraphs of *Economy and Society*).

Clearly, the type and structure of subjectively intended meaning can vary widely, depending on the kind of social action in question, just as long as there is some orientation towards another agent's behavior. First of all, action need not be *rational*. Weber distinguishes two kinds of non-rational action. Non-rational action can be either *traditional* action, i.e. action that follows some "blind routine" (*dumpfe Gewöhnung*), or it can be of the *affective* kind. In contrast to traditional action, affective action involves a clear consciousness of the action. Social action, too, can be non-rational, because routine behavior and emotional acts can be oriented towards other agent's behavior. Just as important as the distinction between rational and non-rational forms of action is Weber's distinction between two *types of rationality*. First, there is *value rationality*. Value-rational behavior is "determined by a conscious belief in the intrinsic value of some ethical, aesthetic, religious, or other form of behavior, independent of its prospects of success." Thus value-rational action shares with affective behavior the element of clear consciousness, which is absent in traditional action. By contrast to both forms of non-rational action, however, value-rational action does involve systematic planning (this is what makes value-rational behavior *rational*). The fourth and last action type is *instrumental rationality* (*Zweckrationalität*). "Instrumentally rational behavior is behavior exclusively oriented to means (subjectively) considered adequate to attain goals (subjectively) clearly comprehended" (Weber 1981: 151; Weber's definition of instrumental rationality in *Economy and Society* will be discussed below). Like as value-rational action, instrumental rationality involves consciousness and planning. In contrast to value-rationality, however, instrumentally rational agents do not disregard their prospects of success. Rather, the effects and side-effects (i.e. both the intended and the unintended consequences) are taken into account.

Thus Weber's action theory allows for a considerable variety of action types. Yet there is another side to the matter. If it is true that there is much space for different kinds of action, it is also true that Weber's methodology clearly favors *one* type,

which serves as the paramount case, or the paradigm of action.[4] There is one type of action which is the *fullest* type, and which therefore is action *par excellence*. Concerning the distinction between rational and non-rational forms of behavior, the emphasis is clearly on the former – in spite of Weber's occasional claims to the contrary. The reason for this rationalistic bias is partly methodological. Weber's rationalism (a label which he rejects) is most evident where Weber uses rationality to distinguish the realm of objects and events that can only be *explained* (i.e. the sphere of natural science) from the realm of the proper objects and events for *interpretation* (e.g. WL 67). Such remarks make one think that Weber is not really serious after all about the possibility of genuinely non-rational action (for a certain behavior would not count as an action if it were plainly non-rational), but even where Weber explicitly does allow for irrationality in the domain of action, he clearly favors rational forms of action over irrational ones – both at the methodological level, and at the level of the content of the analysis. It is true that we can "understand even the irrational exertions of the most excessive emotions", but if Weber claims that according to his view, these are "just as accessible" to interpretation as "the chains of rational 'reasoning'" (WL 100), he seems to be slightly exaggerating, to say the least. Weber's methodology clearly favors rational action, because rational behavior is simply more intelligible than non-rational behavior. And within the domain of rational action, it is *instrumental* rationality that "possesses the highest measure of 'self-evidence'" (Weber 1981: 151). Instrumentally rational action simply has the "most understandable kind of meaning structure" (WL 408; 127, a statement which clearly flies into the face of any claim to a non-rationalist hermeneutics). Thus it is no coincidence that instrumental rationality is the *default* action type that is presupposed in Weber's famous methodological tool, the *ideal type*.

It appears that both the critical and the defensive view of Weber's action theory have their point. It is true that Weber does allow for a rich variety of action types. But it is also true that he methodologically narrows his focus on one particular kind of action, namely action of the instrumentally rational kind. On this line, it seems that Weber believes instrumentally rational action as the "fullest" action type, the other types being more (traditional and affective action) or less (value-rational action) deficient.

The primacy of instrumental rationality over value rationality is particularly evident from Weber's remarks on communal action in his essay *On Some Categories*. Weber does allow for the possibility of value-rational communal action. But the instrumentally rational case is just as paradigmatic for the structure of meaning of communal action as it is for the meaning of any action. Communal action that is motivated in "ideal values", and that is oriented to norms of conduct or to the fulfillment of duties without there being any calculation of the consequences involved, does occur, but Weber considers this as just a "limiting case". In the "normal case", Weber says explicitly, communal action implies a degree of *instrumental*

[4] It is not surprising that, in the received literature, these two tendencies separate Weber's critics from his defenders. For the latter see, e.g., Baurmann (1996: 283ff.).

orientation.[5] To the degree that communal action is instrumentally rational, the agent does not simply do what she thinks she *owes* to the other (in terms of duty, or in terms of any other ideal values such as good taste). Rather, the decisive element of communal action is the following. The agent takes the other agent's *expected behavior* into account in calculating the best means to realize her goals.

> An important [...] normal component of communal action is its meaningful orientation to the *expectations* of certain behavior on the part of others and, in accordance with that, orientation to the (subjectively) assessed probabilities for the success of one's own action. (1981: 159)

This is parallel to the case of non-communal (or, in the terminology of *Economy and Society*, non-social) action. Here, the agent takes into account the expected restrictions and the given circumstances of the objective surrounding world while calculating the best means to his or her end. The only difference between non-communal and communal instrumentally rational action is that, in the latter case, the circumstances that have to be taken into account happen to be other agents' behavior. Significantly, Weber's definition of instrumental rationality at the beginning of *Economy and Society* covers both cases. A certain behavior is instrumentally rational if it is determined by the agents' "expectations of the behavior of objects of the external world and of other human beings", with "these expectations serving as 'conditions' or 'means' for rationally pursued, weighed goals" ([1921] 1980: 12; hereafter quoted as WG). Instrumentally rational communal agents simply do whatever is best to realize their goals (weighing the effects against the expected side-effects), *given the expected behavior of the other agents*. The basic structure of expectation is, as Weber says explicitly, "basically (...) the same" whether B is an inanimate object, or whether he is another agent.

At the same time, however, Weber sees that there is one decisive difference involved here, which captures nicely the modern distinction between paradigmatic and strategic rationality. Let me quote the decisive passage from Weber's *On Some Categories:*

> A subjectively rational agent can also base his expectations of certain behavior from the part of the others on his subjective belief that he can expect subjectively *meaningful* behavior from others, and that he can thus predict, with varying degrees of accuracy, the probabilities arising from certain meaning relationships. (1981: 159)

The obvious difference between the case in which B is a stone and the case in which B is another agent is this: in the latter case, the latter behavior is *action*, i.e. it has *meaning*. Is this just one case among many as the passage seems to suggest? Or is there more to that fact, so that the expectation of B's behavior is not "basically [...] the same", after all, but of an entirely different kind? Let's have a closer look at the matter.

---

[5] This is particularly obvious from the fact that, in the later chapters of his essay On Some Categories, Weber seems to regard a feature that is only characteristic of intstrumentally rational communal action to be a feature of all of communal action, namely the (cognitive) expectation of the other's behavior as a base for the calculation of one's own course of action.

First of all, it seems plausible to ascribe to this case a much more prominent a role for the concepts of social action or indeed communal action than Weber does. If we strictly stick to the letter of Weber's definition in *Economy and Society*, it does not seem necessary for A to conceive of B as an agent for his action to be social (or communal in the terminology of *On Some Categories*), just as long as B *is in fact* an agent. But this does not seem plausible. There is no reason why A, who mistakes B for a stone, and whose action, in its subjectively intended meaning, is oriented to B's behavior, should thereby be a case of social action. The orientation to another agent's behavior that is part of the definition of social action cannot be simply *de facto*; rather, A has to *believe* that B is an agent, just as is implied in the above quotation. (Another question, which will not be addressed here, is whether *belief* is sufficient, or whether *knowledge* is required. What about an agent who mistakes a stone for a person? Is this a case of social action?)

Let's call this the action-orientation condition. I think it is plausible to assume that it has to hold where action is social. For action to be social (or communal), the agent has to orient his or her behavior not just on some other agent's *behavior* (which he might mistake for a natural event). Rather, he or she has to see that the behavior in question is *another agent's*, and that it is *subjectively meaningful* for him or her (i.e. the other agent). In other words, the agent has to take the behavior in question to be *an action*. It seems quite obvious to me that the action-orientation condition conforms to the spirit of Weber's theory, even though it might be the case that, in *Economy and Society* at least, it cannot be found there explicitly.

If this is the case, communal actions are cases in which A orients his behavior to B's expected behavior, which he interprets *as action*. If this is true, however, a problem pops up. To understand the issue at stake here, it is important to keep in mind Weber's abovementioned distinctions, i.e. the distinction between rational and non-rational behavior, the distinction between value-rational and instrumentally rational behavior, and, above all, the distinction between "one-sided" or unilateral social action on the one hand, and social action of the kind of *social relation* on the other. Let's start with the paradigmatic case of social action in Weber's theory: unilateral instrumentally rational action. Here, there seems to be no problem with the above condition that the orientation to another agent's behavior cannot simply be *de facto*. A, who hides behind a tree to ambush B does not expect B's coming by the tree to be a mere natural event (as he expects the apples to fall from the trees in autumn), but an *action*, i.e. as motivated in B's putative aim to go for a walk, or some such. It seems clear that an A who meets the action-orientation condition has certain advantages over an A who sees B's behavior merely as a natural event. The advantage concerns what instrumental rationality is all about: the prospects of success. If A perceives B's behavior *as action*, and grasps the *subjectively intended meaning* of the behavior in question, he will be able to form more reliable predictions and be more successful in forming the corresponding expectations concerning B's behavior. Thus, for the case of unilateral instrumentally rational action, action-orientation is a simple instrumental advantage in terms of goal effectiveness.

What about the case of social relation? What about *reciprocal* instrumentally rational action? If (1) A bases his choice of means on his expectations concerning

B's behavior, and (2) A believes that B is instrumentally rational, too, and if (3) the orientation is (assumed to be) *mutual*, the situation in question meets the criteria of Talcott Parson's concept of *interaction:*

> In interaction ego and alter are each objects of orientation for the other. The basic differences from orientation to nonsocial objects are two. First, since the outcome of ego's action (e.g. success in the attainment of a goal) is contingent on alter's reaction to what ego does, ego becomes oriented not only to alter's probable overt behavior but also to what ego interprets to be alter's expectations relative to ego's behavior, since ego expects that alter's expectations will influence alter's behavior. Second, [. . .] this orientation to the expectations of the other is reciprocal or complementary. (Parsons and Shils 1959: 105)

Parsons (1951: 10) labels the implicit structure of *interdependent expectations* with the term *double contingency* (or *multiple contingency*). This structure is most important for the development of Parsons' thought. The idea of double contingency is conspicuously absent in Parsons' earlier work, especially in his *Structure of Social Action* ([1937] 1949) even though, in his interpretation of Weber's action theory, Parsons had all the necessary means to see this structure. It seems obvious that there is a connection between Parsons' becoming aware of the structure of interaction, and his later turn to systems theory. Indeed, his turn away from action theory and towards systems theory seems to be *motivated* by the view that it is simply impossible to fully accommodate the structure of interaction in an action theoretical conceptual framework (as we shall see, this line of reasoning is particularly obvious in the case of Luhmann's systems theory).

But why should double contingency be a problem for action theory – and an insoluble one at that? The following interpretation of the problem imposes itself. If A is instrumentally rational, he will have to base his decision over the choice of available means (behavior) on an expectation concerning B's behavior. *If* he expects B to be calling back, he should wait, if he doesn't, A should call B himself. This, however leads into an *infinite regress* where A believes that B is instrumentally rational, too, and meets the action-orientation condition. A now believes that B bases her choice of means (behavior) on his (B's) expectation of A's behavior. The expectations are, in other words, *mutually interdependent*, and cannot serve as a basis for the choice of means. The issue at stake here is currently being discussed in Rational Choice Theory, where the focus is on the question of how agents can coordinate their choices rationally.[6] In Weber's terms, the problem is that, in such situations, A's orientation to B's "subjectively meaningful" behavior does not lead to an expectation of probable courses of actions on which A can base his choice of means. If according to Weber's definition, instrumental rationality requires A to base his choice of means on his expectation concerning B's choice, this presupposes that A's expectation and his choice are independent from each other. This, however, is not the case where A assumes that B's choice will be based on B's expectation concerning A's choice, and where A assumes that there is a mutual belief that this is the case (this need not be *common knowledge*). Thus it is obvious that, to the agents,

[6] Cf. Schmid (2007), Chapter 6 in this volume.

this double contingency (or multiple contingency) poses an insoluble "deconditionalizing problem." There is simply no firm ground on which instrumentally rational agents can base their calculations of the best means to realize their goals. Instead of leading to better predictions, and more accurate expectations (as in the one-sided case mentioned above), agency-orientation here leads instrumentally rational agents into total paralysis. Instead of coming up with any prediction of B's likely behavior, A gets lost in an infinite regress, forming an expectation of B's expectations of A's expectations of B's expectations, etc. pp. The question is: how can this consequence be avoided? How can, in other words, the problem of strategic interdependence be solved?

§41 Consensus

In all of Weber's methodological work, there seems to be no proof that he ever became aware of the full extent of the problem of interdependent expectation and its consequence for his theory of instrumental rationality. Indeed, Weber has often been criticized for not addressing the problem of strategic interdependence of decisions at all.[7] In the parlance of today, Weber's theory of rationality in action got stuck on the level of *parametric rationality*.[8] But there are some passages that can easily be read as indicating some degree of awareness of the issue at stake here. The most striking example is a statement in the essay *On Some Categories*. Weber here talks about the role of expectations in instrumentally rational forms of communal actions, which leads him very close to the problem of strategic rationality. Weber says here of the expectations characteristic of communal action that they are marked by an "absolute instability" (WL 422). Significantly, he continues by saying that communal action needs to be *normatively supported* or *integrated* in order to be stabilized. Talking of the agent, he continues, almost hastily:

> specifically, his expectations may be based on an 'understanding' with another or with others; he then believes that he has reason to expect compliance with the 'agreement,' according to the meaning which he himself attributes to it. This alone is enough to give communal action a specific qualitative particularity, for this significantly enlarges the area of expectations toward which the actor believes he can rationally orient his actions. (Weber 1981: 159–160)

In the light of the abovementioned problem of the strategic interdependence of expectations, this *element of normativity* seems to be *mandatory* for rational interaction to be possible at all, and not just one option among others. If this is true, Weber's claim seems to have far-reaching consequences for the architecture of his theory. If it is true that mutual, instrumentally rational communal action (i.e. an instrumentally rational "social relation") *presupposes* some form of explicit mutual

[7] For a very clear statement of this fact see Norkus (2001).

[8] "The parametrically rational actor treats his environment as constant, whereas the strategically rational actor takes account of the fact that the environment is made up of other actors, and that he is part of their environment, and that they know this, etc." (Elster 1979: 19).

commitment to a normative order, or some form of explicit agreement, to which the agent expects the relevant others to stick, there simply is *no conceptual room left* for this particular action type. For instrumentally rational social action based on explicit agreement comes very close to – and indeed is *identical* with – another one of Weber's categories. It is *associational action (Gesellschaftshandeln)*. In *On Some Categories*, Weber gives the following definition of the term:

> Communal action shall be called 'associational action' when and insofar as (1) it is oriented in meaning toward expectations that are held on the basis of agreements, (2) the formation of these agreements has resulted purely rationally (*zweckrational*) in view of the expected action of the associated persons, and (3) the orientation of meaning is subjectively rational. (Weber 1981: 160)

The problem is this. We have seen earlier that, for Weber, the paradigmatic case of action is instrumentally rational action, that the proper object of social science is communal action, and that the paradigmatic case of communal action is reciprocal. Now, if this is what *associational* action is, Weber's claim that associational action is simply a *special case* of communal action, and that there are other forms of communal action, even paradigmatic ones, that are not of the associational kind, seems to be undermined. To uphold this architecture of his theory, Weber needs to identify cases of communal action that are reciprocal, but not of the associational kind. And indeed this seems to be not only a question of the structure of his theoretical edifice, but required by the "things themselves". Reciprocal communal action *is* more fundamental than associational action, and indeed associational action *presupposes* that there are communal actions of the reciprocal kind. How else if not by means of some reciprocal communal actions should the formation of agreements on which associational action is based ever come about? If there are no social relations, and if the relations are not reciprocal, there cannot be such a thing as an agreement, and therefore no agreed upon social order. Without acting in reciprocal social relations, people cannot enter an agreement, and form a contract, and cannot, therefore, perform associational actions (which presuppose agreements, contracts, or some such social orders). But the extension of the concepts of "agreement" and "reciprocal social relations" are not simply co-extensive. All agreements imply reciprocal social relations (I am in an actual agreement with you precisely insofar as you are in an actual agreement with me), but not all reciprocal social relations are agreements of the explicit contractarian kind that is presupposed in Weber's concept of associational action. Rather, such agreements are reciprocal relations *of a special* (and especially *complex*) *kind*. Therefore, it would be a mistake to approach the structure of reciprocal social relations from the analysis of the structure of agreements. The analysis should run the other way around: we need to understand the structure of reciprocal social relations first, before any progress can be made in the analysis of the structure of agreement.

If this is true, we should hold on to the overall architectural plan of Weber's action theory, at least as far as the relation between communal action and associational action is concerned. Associational action is a subclass of communal action. But if the foregoing conjecture is true, if we need a concept of social *normativity* to understand how communal action can be reciprocal, and if that normativity cannot

be pulled out of the hat of agreement (because agreement presupposes reciprocal communal actions), we are left with the following question: what kind of action can there be that is normatively stabilized (thereby avoiding the problem of the "absolute instability" of interdependent expectations), yet not be based on agreement?

This is the point where we finally come to the topic of this chapter. For this is precisely the role of *consensus* and *consensual action* in Weber's theory. *Consensus* is the missing link between communal action and associational action.[9] The term *Einverständnis* (consensus) describes the fact that agents can form reciprocal expectations without getting lost in some infinite circle or loop of interdependent expectations, because agents can be linked by some form of *mutual committment*, without there being agreement or some explicit normative social order involved. It is characteristic for Weber's concept of consensus, and indeed one of its advantages, that *consensus is not based on agreement*. Weber shows clear awareness of the fact that consensus, if the term is to play its structural role in the architecture of the theory, cannot be understood as some "tacit agreement", either. The reason is simply that consensus is whatever makes agreements, including tacit ones, normatively binding (cf. WL 433), and not the other way around. Thus consensus is the source of all social normativity, of institutionalized forms (contracts) as well as of other forms. Weber defines "consensus" in the following words:

> ['Consensus' is] the fact that an action oriented on expectations concerning the behavior of others has an empirically realistic chance of seeing these expectations fulfilled because of the objective probability that these others will, in reality, treat those expectations as meaningful and 'valid' for their behavior, despite the absence of an explicit agreement. It is conceptually immaterial which motives underlie these expectations about this behavior of others. Communal action insofar as it is oriented on such probabilities of 'consensus' shall be called 'consensual action. (Weber 1981: 186, and also WL 432)

Against the background of the above considerations, this definition raises a whole series of questions. First and foremost, the normative element needs to be questioned. At first sight it seems that the element of "validity" concerns B's expectations only, and does not entangle A. If there is consensus, B takes A's expectation to be of some normatively binding quality to him. Yet Weber continues by emphasizing that the reason why A should expect B to feel somehow normatively bound by A's expectation are "immaterial", thereby suggesting that the point is not that B really has to treat A's expectations as "valid" for there to be consensus, but that A has to *believe* that B does so. For there to be consensus, it is not enough that (1) A expects B to behave in a certain way, and that (2) B treats A's expectation as normatively binding. Rather, (3) A has to *believe that B treats his expectation as normatively*

[9] It is true that, in his essay, *On Some Categories*, the section devoted to *consensus* (which fills almost half of the entire essay!) comes only after the section on associational action (which in turn follows the analysis of communal action). But it is clear that, within the logic of Weber's action theory, consensual action occupies the middle position. Indeed this is the succession of terms when Weber enumerates the types of action elsewhere (cf., e.g., WG 381): "communal action, consensual action, and associational action". This mirrors the foundational structure: association presupposes consensus, and consensus, in turn, presupposes communal action.

binding (for whatever reasons A believes B to have for this) and that B conforms to A's expectation "in reality" *because* he does so. This not only seems to be Weber's view; it is also more plausible, for it excludes the case in which A expects B to conform to A's expectations for any other reasons. Consider the case in which A expects B to do x because A thinks x maximizes B's pleasure, while B knows that A expects him to do x and does x because he believes that he *should* conform to A's expectations. This should not be considered a case of consensus for two reasons. First, it does not seem plausible to call any of A's action that are based on such expectations *consensual* because this is counterintuitive given the ordinary language meaning of the word. Second, and much more importantly, such cases do not explain how the problem of interdependent expectations can be solved, because for this to be the case the normative element has to be part of *both agents'* "subjectively intended meaning".

If it is only *accidental* that B conforms to A's expectation for normative reasons, and not in any way *expected* by A, it seems plausible to call his action *social*, but not in any way consensual. For pure social action, it is immaterial what motifs agents ascribe to each other. If, however, the general aim of Weber's methodology is to construe the essential categories of action theory in terms of *subjectively intended meaning*, Weber cannot use "objective probabilities" of behavior to distinguish social action from consensual action; rather, the *differentia specifica* between consensual action and other kinds of social action has to be a matter of the *content* of the agent's "subjectively intended meaning", i.e. of what he intends and what he believes. Thus consensual action presupposes that A takes B to fulfill A's expectations for the reason that B thinks that he *should* do so. In the reciprocal case, he will not only have to take B to assume that he, A, is normatively bound, too. Above all, he has to take *himself* to be normatively bound to conform to B's expectations.

It seems that the kind of expectations at stake here defines the concept of a social norm. According to the simplest definition, a social norm is a special type of social regularity. A social regularity is a reason to expect that people will exhibit a certain type of behavior. If the regularity in question is of the special kind of a social norm, there is a reason to believe that people will exhibit the type of behavior in question *because they believe they can* (*justifiably*) *be expected to do so*.

If this reading of the structure of consensual action is right, it seems, however, that there is still something deeply wrong about the architecture of Weber's theory. The problem is that the characteristic feature of consensual action seems to be more than just a simple *differentia specifica* of social or communal action.

Upon closer consideration, it appears that Weber's talk of expectations in the definition of consensual action quoted above, and the use of the word in the definition of social or communal action, covers up an ambiguity in that term. The difference at stake here is between expectations of the *cognitive* kind, on the one hand, and *normative* expectations, on the other. While cognitive expectations are part and parcel of the concept of social action (as the foremost way in which agents take into account other agent's behavior), it might seem that the expectations in the above definitions are of the *normative kind*. Thus the question whether or not consensual action fits into the general theory of social action depends on the relation between these two kinds of expectations.

Here are some differences between the two. Most obviously, the *direction of fit* is different. In the case of cognitive expectations, the direction of fit is mind-to-world; in the case of normative expectations, it is world-to-mind. The difference between the two directions of fit becomes particularly obvious in the case in which the expectations are not fulfilled. If A *cognitively* expects B to perform some action, and if it turns out that B does not perform the respective action, A will put the blame on himself rather than on B. A sees now that he simply *miscalculated* B's behavior, i.e. that his expectation was *mistaken*, which is his problem, not B's. If, however, A *normatively* expects B to perform an action which he or she does not perform, A will put the blame on B rather than on himself: it's not that A was *mistaken* in expecting B's action; rather, *it is B's fault not to do what he or she was expected to do*. In the first (i.e. the cognitive) case, A shouldn't have expected B's action; in the latter (i.e. the normative) case, B shouldn't have failed to do what was expected from him or her.

This is the reason why the latter type of expectation is generally much more resistant to contradictions with experience than the first. Cognitive expectators are ready to *learn*, i.e. to adapt the content of their expectations to what they know about the structure of the world (hence the mind-to-world direction of fit). One cannot cognitively expect one's pet to be housebroken if one believes it is not. But one can *normatively* expect one's pet to be housebroken if one thinks that this is what the pet *should* be. *Normative* expectators are more likely to *teach* than to learn, to change the world rather than their expectations, i.e. to see to it that the empirically observed behavior in question meets their requirements (hence the world-to-mind direction of fit).

Another way to approach this difference between cognitive and normative expectations focuses on a difference in the kind of *intersubjective relations* involved in each of these cases. The relation between the subject of the expectation, and the person whose behavior is expected is fundamentally different: cognitive expectators *take other people's behavior into account*, while normative expectators *count on other people*. This difference is fundamental indeed, and it is obscured not only by the ambiguity of such words as "expectation", but of other words such as the verb "to rely on", too. If we climb up a big tree, we rely on the strength of the branches in a sense that is fundamentally different from the sense in which we rely on our comrade while climbing on a rope in the mountains, and it may well be that our last words will mark this difference if our expectations should be disappointed.

The question to be answered here is the following. If the expectations involved in consensual action are of the normative rather than of the cognitive kind, does that mean that Weber can assign this type of coordination its place within the framework of his theory only at the price of an equivocal terminology? Does Weber use the double meaning of the word "expectation" for the purpose of pulling consensus out of the hat of social action?

This is precisely what both of the opposing camps in German social theory seem to think. In the following, we shall have a closer look at how Niklas Luhmann and Jürgen Habermas try to solve Weber's problem, and how their choice of basic conceptual framework is determined by their attempt to succeed at this task. Parsons

and Luhmann think that this cannot be done within an action theoretic framework, and that it is necessary to turn to systems theory in order to get a grip on the phenomenon. Habermas, in turn, sticks to action theory, but switches from intentionalism to a linguistic foundation for social theory. I believe that neither Luhmann's nor Habermas' solution really works, and I will try to show why. In the last section, I shall make a suggestion as to how Weber's problem could be solved within action theory, and within an intentionalist setting at that. All that is needed is a theory of *collective* intentionality.

§42 Consensus and Contingency

For the later Parsons, the problem of double contingency and the role of social norms in interaction ultimately show that social theory cannot be based on an account of (rational) action. The action theoretic focus on "subjectively intended meaning" is simply too narrow to capture the structure of interaction, because within this framework, it cannot be explained how agents come to have normative expectations. Consensual actions imply some kind of normative order: that normative order is necessary to solve the problem of the circle of interdependent expectations, and to counteract the "absolute instability" (Weber) of strategic interdependence.[10] In other words, action theory cannot explain, but indeed *presupposes* the existence of a normative order, at least if it extends to actions of the consensual kind. Since there is no doubt that interaction is in fact possible, and since this cannot be explained within an action theoretic framework, the structure of interaction cannot be captured within an account that is focused on the "subjectively intended meaning" of behavior. Interaction is not a matter of isolated individual *unit acts* (Parsons' term). The actions which interacting agents perform have to be taken as elements *within a system*.[11] Interaction requires a consensual element, and that, in turn, presupposes the existence of a generalized normative order, i.e. a shared system of symbolic meaning, conventions, and other cultural standards. Because these cannot be reduced to the "subjectively intended meaning" of the participating individual agents, the theory of social action has to go beyond action theory. Consensus cannot be understood from the perspective of the subjectively intended meaning of the agents; rather, the analysis has to account for the structural conditions of possibilities for consensual action. And this is precisely what systems theory is all about. This radical shift in

[10] "The most important single condition of the integration of an interaction system is a shared basis of normative order. Because it must operate to control the disruptive potentialities (for the system of reference) of the autonomy of units [...] such a basis of order must be normative" (Parsons 1968: 439).

[11] This is already expressed in Parsons ([1937] 1949: 740). This is an early indication of the line of development James Coleman criticized in the following words: "Parsons abandoned his attempt to found social theory in a theory of rational action; he reverted to classification schemes that were no less sterile in his hands than in the hands of those he criticized" (Coleman 1992: 49). For Coleman's own treatment of the problem of double contingency cf. Coleman (1990: 901ff.).

perspective entails a complete reversal of the order of analysis. Instead of giving an account of social facts based on the interpretation of the subjectively intended meaning of the agents (as it is done in action theory), social facts are now, in turn, seen as based in the "mutuality of socially structured relationship patterns," (Parsons 1954: 359) and these are taken to be the *preconditions for the possibility* of such individual mental states, which have to be listed and classified in order to develop an adequate account of interaction. Thus Parsons' use of the argument that social norms are a condition for the possibility of interaction is such that the action theoretical, hermeneutic approach is replaced by an analytical, classificatory reconstruction of systemic structures from an external point of view.

To the critics, this radical shift of perspective clearly shows that systems theory does not *solve* the problem, but simply discards the point of view from which it is a problem. The perspective is shifted from the search for *reasons for action* that are fit candidates to rationalize normative expectations to the search for the structural preconditions for empirical motivational patterns. Of course, Parsons' underlying thesis is that the structure of the meaning of consensual action simply goes beyond the subjective intention, i.e. beyond the perspective of the participating agents, and is, therefore, of systemic nature. According to Parsons, the cultural symbols (which include normative components) are responsible for the success of coordination, and thus have to be taken into account even though they are more like *causes* than *reasons* for consensual action. Loosely speaking, it is the system that is in control of the situation, not the agents themselves, "the system being so *geared into the action system* of both ego and alter that the external symbols *bring forth* the same or a complementary pattern of orientation in both of them. Such a system of normative orientation is logically the most elementary form of culture" (Parsons et al. 1959: 16; my emphasis). As Parsons continues, this "internalization of culture patterns" *creates* "personality" (ibid.: 22) as part of the social system, which in turn means that Parsons conceives of personality entirely in terms of cultural conventions. In a sense, the system replaces the individual as *the agent* of consensual action. This might seem implausible in itself, yet there is another consequence which seems particularly hard to swallow. The view that social agents are "steered" by social norms is at the heart of what has rightly been criticized as Parsons' "oversocialized" account of action, leading to Parsons' notorious conventionalism. From the point of view of systems theory, *all* action seems to be norm-regulated behavior, with norm being somehow the more fundamental category than action in the order of the concepts. The phenomena of deviance and dissidence, conflict and innovation are excluded from the analysis on the conceptual level (cf. Wrong 1992: esp. p. 216).

This is an immediate effect of the way in which Parsons tries to solve the problem of Weber's theory by reformulating the normative component of consensus as "normative social order" in a systems theoretic perspective. And even if we leave aside the point that Parsons' systems theory *avoids* Weber's problem rather than *solving* it, and that the way in which the system replaces the agency as the center of control seems somewhat questionable, these consequences seem much too difficult to accept calling Parson's way of dealing with Weber's problem successful.

It is interesting to see how Niklas Luhmann, in his version of systems theory, addresses this problem. Luhmann tries to overcome Parsons' conventionalism, which is a consequence of Parsons' turn to systems theory, with his own version of systems theory. The third chapter of *Social Systems*, Luhmann's *opus magnum*, is entitled *double contingency* (Luhman 1984: 148ff.). It starts with a critical discussion of Parsons. Luhmann here argues that Parsons' mistake was to introduce his concept of normative order as some kind of *compensation* for the problem of double contingency, i.e. as some external element that makes double contingency *disappear*, instead of seeing double contingency as a conceptual ingredient, and indeed as a precondition for the possibility of any communication. In other words, Luhmann seems to think that Parsons' "oversocialized" view is caused by the half-hearted way in which he dealt with double contingency, and by the quest for intersubjective "convergence" of meaning. With a more courageous shift to systems theory, Luhmann claims, all of that is left behind. Consensus is a problem only for those who think that there really *is* such a thing as intersubjective identity of meaning. By renewing systems theory on the base of the concept of *autopoiesis*, Luhmann claims that he can finally liberate social theory of all older claims to consensus and intersubjective convergence. The mistake of all older theories, Luhmann claims, is to start out with the category of meaning, which is relative to a subject or system, but then to claim that different subjects and systems can somehow converge, and to construct meaning as an inter-subjective category. According to Luhmann, this attempt is simply futile, and should be discontinued. If meaning is relative to one system of subject – as Luhmann, too, thinks it is – it is relative, full stop. Thus the theory of autopoietic systems is aimed at taking the ontological subjectivity of meaning seriously. The modern experience that meaning can vary, that there is no inter-subjective or inter-systemic unity that guarantees some prestabilized harmony, some convergence, let alone some universal identity, is not taken as a sign of crisis which has to be overcome, but as a fundamental feature of social reality which we should finally take for what it is: everything could always be different. There is no place for sure bets in social reality, not even in what might appear to be the firmest consensus. Thus in his reading of double contingency, Luhmann stresses the modal feature of the word contingency:

> Contingency means that being depends on selection which, in turn, implies the possibility of not being and the being of other possibilities. A fact is contingent when seen as selection from other possibilities which remain in some sense possibilities despite a selection. (Luhman 1976: p. 509)

The fact that, in the situation of strategic interdependence, individuals cannot base their expectations on each other's choices, i.e. that the attempt to base one's own choice on the other's expected choice leads into an infinite regress of interdependent expectations, does not mean, according to Luhmann, that some external normative element (normative social order) has to be there so that interaction can start. The fact that individuals are "black boxes" to each other, the "darkness of mutual intransparency", as Luhmann puts it in this chapter, is nothing that has to disappear for communication to become possible. The whole point of Luhmann's theory is to drop the idea of the convergence and mutual transparency of the meaning of

behavior. Mutual opacity and the relativity, indeed, the *privacy*, of meaning is not a problem which has to be overcome for the social system to be possible; rather, these are *preconditions for the possibility of the social system*. The social system does not emerge *in spite*, but *because* of double contingency. The element of commonality which Parsons declared to be a condition for social action, and which he found in the normative social order, is simply declared inessential and consequently dropped. The social does not have to be anchored within the realm of the intended meaning of the participating individuals. In Luhmann's view, the individuals are – and remain – utterly opaque to each other. The social does not resolve double contingency; rather, it emerges from *whatever happens* in situations of double contingency. This is what communication is, according to Luhmann: whatever happens in situations of double contingency. Thus double contingency defines and underlies the social. And because the social is based in double contingency, it is not to be reduced to the participating individuals (individual systems), but rather an emergent level of system.[12]

> For a social system [. . .] it is not necessary that the systems which are in double contingency to each other can see through each other, and prognosticate. The social system is a system for the very reason that there is no basic certainties of state, and no predictions relying on these. (Luhman 1984: 157)

Luhmann defines the elements of the social system as *communication*. Communication, just like the expectations of the participating individuals, is a matter of *meaning*. But, as is apparent from the above considerations, the term "meaning" is rather equivocal in Luhmann's theory. Meaning in terms of communication is not meaning in terms of the intentions of the participating individuals. Quite to the contrary, Luhmann's way of dealing with the concept of double contingency implies the radical distinction, the categorical rift between the two kinds of meaning, for which his theory is justifiably infamous: the distinction between communication (i.e. meaning at the level of the social system) and consciousness (i.e. meaning at the level of the thoughts of the participating "psychic systems"). Luhmann rejects any attempt to go from meaning in terms of whatever is "subjectively intended" to meaning in terms of communication. In a kind of a bold strike through the Gordian Knot of double contingency, the meaning of meaning is simply cut in half: meaning

[12] The decisive passage in *Social Systems* reads as follows: "The black boxes generate whiteness, as it were, when they meet, or at least sufficient transparency to deal with each other. By means of *their mere hypothesizing* they generate certainty of reality, because that hypothetizing leads to the hypothesis of hypothetizing from the part of *alter ego.* The assimilation of meaning materials to this level of order presupposes [. . .] two self-referential systems, which observe each other. For those few aspects which are of importance for their dealings, their capacity for information processing might be sufficient. They remain separate, they do not merge, they do not understand each other better than before. They concentrate on what they can observe about the other *qua* 'system-in-an-environment' in terms of input and output, and they learn self-referentially, within their own observer perspective. They can try to influence whatever they observe by means of their own action, and they can learn from the feedback they get. In this way, an emergent order can come into existence, an order that is dependent on the complexity of the systems that make it possible, but that does not depend on the possibility that this complexity can be calculated, or controlled. We call this emergent order the social system" (Luhmann 1984: 156f.).

is the medium of the integration of psychic systems as well as of social systems. But "social meaning" (in terms of communication) cannot be derived from "psychic meaning" (in terms of the meaning of the thoughts of individuals), and neither can the content of thoughts be derived from communication. Thus communication is not shared meaning, but something entirely different.

From the point of view of the architecture of the theory, this move might seem fascinating. From an analytical perspective, however, this is so utterly implausible that it seems difficult to understand why Luhmann's systems theory has been successful for so long. Indeed, the distinction seems rather self-defeating. Why should we *think* there could be anything to the things which Luhmann *tells* us if there is no bridge between what is told and what is thought?

Granted, the two do not always converge. In an introduction to Luhmann's systems theory that is quite widely used in the German speaking world, the consequences of Luhmann's categorical rift between thought and communication is illustrated with the example of a physician's wandering thoughts during an interview with his patient. In cases of routine conversation, such phenomena as talking on auto pilot might be quite frequent indeed, but it is simply nonsensical to declare this the normal case. Even if they are wrong more often than they think, people say what they say because they assume they say what they think, i.e. that the meaning of communication reflects the meaning of the speakers' thoughts.

This is very basic folk psychology and common sense indeed, and to go against it would require better reasons than finding an elegant solution to the problem of double contingency. Not surprisingly for a social theorist, Luhmann himself is not willing to stick to his own claim, when his concern is not with the conceptual aspects of his grand theory but with its descriptive content. More than anywhere in the gargantuan body of Luhmann's work, this is the case in his writings on education science. Analyzing the relation between the student's thoughts and the communication going on in the classroom, Luhmann departs from his idea of an unbridgeable difference between thought and communication, and instead claims that there is some "congruence (. . .) of psychic and social events" (Luhmann 1987: 179). Luhmann does not explain any further what he means by "congruence", but it seems quite clear that he does not do so because, whatever congruence might turn out to be, it will hardly be compatible with his clear conceptual cut between thought and communication.

Thus Luhmann's own analysis proves his "solution" to the initial problem wrong. We are still stuck with the task of finding a way to show how interaction and "subjectively intended meaning" relate to each other, i.e. of what reasons agents might have for normatively expecting other agent's behavior. Luhmann's bold stroke through the Gordian Knot does not make the problem disappear. The trouble with Weber's category of consensus remains. Many of the complexes of behavior that are of interest to social science are based on the fact that individuals do count on each other, and can successfully do so. In these cases, the "expectations concerning the behavior of others has an empirically realistic chance of seeing these expectations fulfilled because of the objective probability that these others will, in reality, treat those expectations as meaningful and 'valid' for their behavior, despite the absence of an

explicit agreement" (Weber 1981: 168; WL 432). An essential feature of these cases is the following. By contrast to Luhmann's claim, these agents do not see each other simply as "black boxes", or as "opaque systems", and do not limit themselves to observe their input and their output. The relation between such individuals (or systems) is of an entirely different stamp indeed. These individuals (or "psychic systems", if you wish) are related to each other by *normative* expectations; they do not just *take each other into account* as elements of the surrounding world; they *count on each other*. This phenomenon (which Weber labels "consensus") cannot be accounted for within a conception that dichotomizes the sphere of meaning into "psychic" and "social". Thoughts and communications are different in many respects. But their relation is more intricate than Luhmann believes. In trying to avoid the problematic consequences of Parsons' attempt to do justice to the role of social normativity in interaction, Luhmann lets himself be carried away into a conception that not only seems rather self-defeating and to fly into the face of common sense. Above all, his conception ends up in a straight denial of the phenomenon which Weber's category of consensus is supposed to capture. In his view, systems always "remain separate, they do not merge, they do not understand each other better than before. They concentrate on what they can observe about the other *qua* 'system-in-an-environment' in terms of input and output, and they learn self-referentially, within their own observer perspective. They can try to influence whatever they observe by means of their own action, and they can learn from the feedback they get." (Luhmann 1984: 156f.)

In other words, systems (psychic or other) do not have normative expectations concerning other system's behavior, because all they do is observe their input and their output, which – if there are any regularities to be observed – leads them to cognitive expectations. But, in reality, there *are* normative expectations. To learn more about their nature, role, and structure, Luhmann's systems theory is obviously the wrong place to turn.

§43 Consensus and Language

Let's now have a look at the other of the two camps in which German social theory has been divided over the last 2 decades of the twentieth century, and see how Weber's problem is dealt with there. For Jürgen Habermas, Weber's essay *On Some Categories* proves that there is an *unofficial version* of Weber's action theory, a version centered on the concept of *consensus*, hidden behind the official version. The official version is, according to Habermas, marked by a concept of meaning which is oriented towards the paradigm of the solitary instrumental action of a single subject. This concept of meaning stands for everything Habermas thinks we have to leave behind in order to proceed to an adequate conceptual framework for social theory. First, the official concept of meaning is *intentionalistic* instead of *linguistic*: it captures meaning in terms of whatever is "subjectively intended" by the agent, and tries to derive the structure of consensual action from the "subjectively intended

meaning" of the agents. Second, Weber's concept of action is *instrumentalist* rather than *norm-oriented*. On the "official" line, Weber describes action primarily in terms of means and goals, i.e. as a matter of the pursuit of instrumental success rather than in terms of social propriety, conformity, or deviance. And third, Weber's account of action is *monological* rather than *communicative*. Action appears as a matter of single individuals who act in light of their own goals, rather than as a matter of group members who act on a shared understanding of their situation.

Thus his "official" Weber nicely epitomizes Habermas' opponents: *Bewußtseinsphilosophie*, intentionalism, functionalism, "monologism". Against all of that, Habermas recommends no less than a radical shift of paradigm. This is achieved with his new "linguistic foundation of sociology." Meaning, Habermas claims, should be seen as a matter of language rather than as a matter of intentional mental states. Following a great many philosophers of his generation, Habermas argues that linguistic meaning cannot be derived from the intentionality of the mental states of the speakers. Rather, the intentionality of mental states is *derived* from linguistic meaning. Thus, in his view, meaning is no matter of "subjective intention", as Weber has it, and is thus not only *contingently* social (insofar as others are in the content of these mental states). Rather, meaning is a matter of shared linguistic practices, and hence *a priori* social (cf. Habermas 2001).

Later on in Habermas' work, however, it becomes clear that Weber plays a *dual role* within the project of the *Theory of Communicative Action*. On the one hand, Weber still serves as the epitome of everything that has to be left behind. On the other hand, however, Weber's essay *On Some Categories* serves Habermas as evidence for his claim that, deep inside the "official" Weber, there is a little "unofficial" Weber struggling to get out. Needless to say that this unofficial Weber is much more to Habermas' taste, since he almost seems to preempt Habermas' own shift of paradigm towards a linguistic foundation of social theory. Following is what Habermas says about Weber's "unofficial" views.

Whereas the "official" Weber tries to conceive of human coordination mostly in terms of the agents' *interests* (*Koordination durch Interessenlage*), the author of the essay *On Some Categories* contrasts this mechanism with an entirely different type of coordination: *coordination by consensus (Koordination durch Einverständnis)*. It is not difficult to recognize that the distinction between these two types of coordination anticipates Habermas' later fundamental distinctions between *System* and *Life-world*, or between instrumental and communicative action.

In Habermas' reading of the unofficial Weber, the distinction between communal and associational action is *within* the domain of coordination by consensus. These action types are, according to Habermas, sub-classes of consensual action: communal action is consensual action with a low degree of rationalization, while associational action is highly rationalized, with rationalization meaning value-rationalization rather than instrumental rationalization. Habermas then reads Lawrence Kohlberg's dichotomy of conventional and post-conventional moral reasoning right into that distinction. The result is the following. In the case of communal action, agents can count on other agent's behavior insofar as their expectations conform to the conventional standards of social propriety. In the case of

associational action, the normative expectations of the participating agents are based on the formal principles of fairness.

As a consequence of Habermas' reading, consensus is conceptually cut off from the sphere of instrumental action. Thus this reading is in a sharp contrast to another interpretation that seems to be imposed by the general architectural structure of Weber's theory, i.e. the attempt to find the sources of consensus and of the social normativity which is presupposed in consensual action somewhere *within* the domain of instrumental action. In the larger context of Habermas' general venture, the *Theory of Communicative Action*, this point is by no means a marginal one. Immediately following (and indeed based on) this somewhat idiosyncratic reading of Weber's categories, Habermas introduces the most important of his own conceptual tools, i.e. the distinction between *communicative* and *strategic action:*

> We call an action [. . .] *strategic* if we consider it under the aspect of compliance to the rules of rational choice and if we assess the degree of efficacy of the influence on the decision of a rational opponent. By contrast, I speak of *communicative* actions, if the plans for action of the participating agents are coordinated *via* acts of communication rather than *via* egocentric calculations of success. In communicative actions, the participants are not primarily oriented towards their own success; they pursue their individual goal under the condition that they can tune in their plans for action on each other on the base of shared definitions of the situation. (Habermas 1981: 385)

In brief, and put very simply, communicative action *is* Weber's consensual action. The special flair of Habermas' reading of the normative element in Weber's concept of consensual action stems from the fact that Habermas claims the element of communication within consensus to be *of a linguistic nature*. Consensus is ultimately the "telos of human language", and to be achieved *only* by linguistic means: "The concepts of speaking and of agreement (*Verständigung*) mutually interpret each other." (Habermas 1981: 387) Thus there is, according to Habermas, a necessary link between the meaning of consensual action and the linguistic competence of the respective agents. As Habermas puts it, a consensus is necessarily a proposition that is mutually accepted. Thus consensus has a "linguistic structure" (ibid.: 386).

I do not think that it is necessary to waste much time with the question of whether or not Habermas is right when he claims that his reading of the "unofficial version" of Weber's theory is "well supported" by Weber's essay *On Some Categories*, as far as interpretive correctness is concerned. It is simply too obvious that Weber does not have anything like Habermas' distinction in mind here, but aims at integrating consensual action into the theory of communal action in just the way as he does elsewhere in his work. In contrast to what Habermas claims, communal action is introduced as the *general term* that encompasses both consensual action and associational action. This makes it utterly implausible, from an interpretative standpoint, to open up a conceptual rift between goal-oriented action and consensus in Habermas' sense. But let's leave these tedious questions of interpretative correctness behind and turn to an issue that is of much greater relevance: is Habermas' concept of consensus *qua* theory of the source of social normativity consistent and analytically adequate?

The main problem I wish to address is the necessary link between speech and consensus as postulated by Habermas. This seems problematic for two reasons. First and foremost, this theory *a priori* limits consensual action to linguistic practitioners, which does not seem particularly convincing considering everyday examples in which consensual action seems to be oriented towards non-linguistic agents. And second, it seems that we need some concept of consensus to analyze the preconditions that enable pre-linguistic beings to enter linguistic communication. These beings have to *see each other* in a certain way, and it seems that the concept of commitment nicely captures some necessary features of their mutual relation. If consensus *is itself* a matter of speech, however, this concept is useless for the analysis of those preconditions of speech.

Let us have a closer look at Habermas' claim first. Of course, Habermas does not want to suggest that each and every consensus has to be a matter of explicit linguistic communication. There is what Habermas calls *lebensweltliches Vorverständigtsein* – a kind of consensus that is based in the meanings of the life-world, which agents simply take for granted in their every day dealings. But, in Habermas' view, the only reason why agents take those tacit life-worldly background assumptions to be consensual is that they *think that possible dissent could be negotiated by means of entering into a discourse*, *which is a linguistic matter*. In other words: if A tacitly assumes that B treats her expectation as valid for his own behavior, she does so because she thinks that, if B were to disagree about the validity of her expectation, they could have a discussion about this and find their way back to a consensus. Thus, in this sense, even those consensual meanings which agents do not talk about, and simply take for granted, are *constituted* by language (in terms of the possibility of a discussion). Consensual agents not only have to assume that their counterparts will accept their expectations as valid; they also have to see each other as possible candidates for an open debate, if the validity of the respective expectations is contested.

This thesis does not seem particularly appealing. Especially if we stick to Weber's definition, it does not seem obvious why consensual action should presuppose the possibility of linguistic communication. Indeed it seems quite frequent empirically that people have normative expectations concerning the behavior of beings, which they *do not* seem to take to be competent linguistic practitioners, or indeed possible candidates for discussions on contested claims to validity. Beings capable of speech are not the only members of the class of actual addressees of normative expectations. Among the beings of whom a certain behavior is normatively expected are those who are not linguistic practitioners *yet* (and perhaps some who are not linguistic practicioners *any more*). And indeed there are some members of this class, which never were – and never will be – competent speakers. This cannot be accommodated in Habermas' theory; without the possibility of linguistic communication, there simply cannot be normative expectations; the only attitude that is possible towards non-linguistic beings is cognitive expectation.

But how do we determine whether or not A expects B's behavior in a normative or in a cognitive way? Even though these kinds of expectation are very different conceptually, as the above considerations have shown, it is difficult to distinguish the two kinds empirically. This is especially so since A's behavior in reaction to

a disappointed expectation does not by itself tell whether his expectation was of the cognitive or of the normative kind. Let's assume that A is a dog owner and B is his pet. If A finds out upon his return home from work that his expectation concerning his pet's being housebroken has been disappointed once again, he may react by drawing the dog's attention to its pile, shout "fie!", and perhaps do some of the other things people usually do in such situations. The point is that from this behavior alone it is not possible to tell which of the following scenarios is true. According to one interpretation of the dog owner's behavior, he is *disappointed* to find out that his expectation was mistaken, because his dog is obviously not yet house-broken after all, and needs some more training. His behavior is an attempt to condition the dog so as not to exhibit that kind of behavior again. According to the other interpretation, however, the dog owner is genuinely *angry* at the dog, because he thinks that his dog did something it *shouldn't* have done, and that he was *justified* to expect house-broken behavior of his dog, and now *punishes* it for its deviant behavior.

These are two very different scenarios from the internal perspectives, even though from an observer's point of view, the behavior is exactly the same. In the first case, the disappointed expectations are of the purely cognitive kind, in the second of the normative kind.

Aside from the epistemological problem of how to determine, from a third person's perspective, if a given case of expectation is of the normative or of the cognitive kind, it does seem obvious that people *do in fact* address normative expectations to beings with whom they could not have a discussion. Of course, such phenomena are per se no counter-evidence to linguistic theories of social normativity such as Habermas'. The mere *fact* that some agent *assumes* herself to be in some kind of consensual relation with a being that is incapable of speech does not proof that non-linguistic consensus is possible. It may well be that consensus is *merely assumed* by the dog owner, which is not quite enough for there to be *actual* consensus. Even according to Weber's notoriously subjectivist definitions, there needs to be more than just an imputation of consensus for there to be genuine consensus: remember that there has to be the "objective possibility" that the expectations of the agent are treated as "valid" by the addressee, too.

There is a difference between merely *believing oneself to be* in a consensual relation with another being, and *actually being* in that kind of relation. It seems to be a necessary precondition for a relation to be consensual that the agents *believe* themselves to be in a consensus. But this is not *sufficient*. There are other criteria to meet. So it might appear that many of the cases in which people assume themselves to be in a consensus with non-linguistic beings are simply *mistaken*, and are not in an *actual* consensus, in spite of what they believe to be the case.[13]

We should not dismiss the possibility, however, that not all cases of imputed consensus with non-linguistic beings are of this erroneous kind. There is another reason to hold on to the idea of a non-discursive normative practice. From a genetic

[13] As an example for this, many dog owners who have normative expectations with respect to their pet's behavior actually tend to mistake their dog for a linguistic practitioner, speaking to their dog as if it were a human being.

perspective, it does not seem easy to see how beings that cannot, in some rudimentary sense, *count on each other*, should ever come to use linguistic means to communicate. These considerations show that Habermas' link between consensus and language is by no means a matter of course. At the same time, it is clear that the idea of a pre-linguistic consensus is not without problems. There are problems on three levels: the epistemological, the moral or practical, and the ontological level. I shall briefly address each of these levels in turn.

From an epistemological point of view, it seems clear that any *verification* that there is an actual consensus presupposes language. Only linguistic practicioners can *confirm* that they are in a consensus. But this epistemological problem does not *per se* disconfirm the possibility of pre-linguistic consensus insofar as it is not inconsistent to assume an entity whose existence cannot be verified.

The second level is more difficult to address. An observation concerning the received literature on the sources of social normativity illustrates the point. There is much to learn from this literature about what entitles agents to have normative expectations that are addressed at the participants of linguistic practices, ranging from Christine Korsgaard's theory of the *Sources of Normativity* to Habermas' Discourse Ethics. In contrast to this, there does not seem to be *one single attempt* even to try to show why having normative expectations concerning the behavior of beings that are not capable of language might be anything else than totally unjustified. And there seems to be good reasons for this linguistic bias in the theory of social normativity. It is obviously rather problematic, if not outright absurd, to have such expectations. Why should somebody be considered to be *obliged* to a certain kind of behavior, if he or she has not committed herself to that norm? And how should such a commitment come about if not in some kind of discursive procedure?

I think that this objection is correct, but only as far as *morally laden* cases of normative expectations are concerned. For a consensus to be morally binding, there has to be consent, and consent implies endorsing a proposition, which is a linguistic matter. But not all normative expectations are morally laden. A dog owner who has normative expectations concerning his dog's behavior does not have to take his dog to behave *immorally* if it leaves its pile on the floor. There are "oughts" of weaker kinds. Let's not enter into a discussion about moral obligations here, and stick to the simpler question of whether there might be reasons to expect other agents' behavior normatively, if these agents are not linguistic practicioners and can therefore not be seen as morally obliged by consent.

Even if we accept these replies to the epistemological and the moral–practical objection, there are huge difficulties left. Many current philosophers claim that it is simply a *mistake* to apply the terms *action* and *practice* to pre-discursive behavior. This suspicion is nourished by current attempts to allow for the possibility of non-discursive normative practices. The most prominent example is Robert B. Brandom's theory. Brandom's explicitly stated aim is to maintain the conceptual possibility of non-discursive normative practices. This, of course, gives rise to the question: whence do those norms come from? Brandom's reply seems to be that they are simply instituted by *sanctioning behavior*. And that, Brandom claims, does not presuppose language. One does not have to use language to teach other people

manners. If necessary, the appropriate use of a stick will do just fine.[14] The most obvious problem of this and other *sanctioning theories of social normativity* is that, for a behavior to be an act of *sanctioning*, the respective norm has already to be in place. A complex of behavior can be sanctioned only if it is measured against a norm. This is what distinguishes acts of sanctioning from mere aggressive behavior. Normative expectations are more fundamental than sanctioning behavior, because the latter presupposes the former. Thus again, there seems to be no place for normative expectations among non-linguistic beings.

Should the idea of pre-linguistic consensus therefore be dropped? I think it should not, and I believe that a closer examination of Weber's problem can show us why. If pre-linguistic consensus is possible, it has to be shown how Weber's problem can be solved without turning either to systems theory, or to a linguistic foundation for social theory. In other words, it has to be shown how consensual action can fit into Weber's framework of goal-oriented action. This is precisely what I will try to do in the remainder of this chapter.

A *caveat* is in order. There are two lines of objection against the following attempt. First, it can be argued that any kind of goal-orientedness presupposes the use of speech, and thus social normativity. Second, it can be argued that an instrumentalist conception of practical reason is *per se* anti-normativist, and cannot accommodate social normativity. Even though the first objection is more fundamental, I will only address the second in the following. In other words, I will simply take the possibility of pre-linguistic instrumental action for granted, and ask the question: how can the concept of consent be accommodated within a theory of instrumental action that does not count the normativity which is constitutive of consensual action among its preconditions? I gladly admit that, for complex calculations of means within multi-layered plans for action, this assumption is highly improbable. For more basic forms of goal-orientedness, however, this assumption might be granted the benefit of the doubt.

§44 Consensus and Commitment

From the above considerations emerges the following picture. The critical discussion of Luhmann's and Habermas' attempts at solving Weber's problem has shown that it is perhaps not entirely implausible to hold on to the following two basic Weberian intuitions: first, we have to give an *action-theoretic* account of consent, i.e. the structure of consent has to be accounted for *from the agents' points of view*, with an eye on the *reasons* those agents might have for whatever expectations they have. Second, consent is more basic than any form of linguistic procedure. Consent is not a genuinely linguistic matter, but rather among the presuppositions of linguistic communication.

[14] Brandom (1997: 201): "I am indeed committed to the possibility of norms implicit in prelinguistic [...] practices. [...] The picture is that what proto-hominids could do before they could talk is to take or treat each other's performances as correct or incorrect by practically sanctioning them."

It is not necessary to assume that Weber's theory of consensus is superior to the theories of his critics in order to think that it might be useful to reconsider Weber's within the context of his own work while trying to find a solution to his problem. The question on which a second look at Weber's conception has to focus is the following: *what precisely is it that makes goal-oriented action such an implausible paradigm for a theory of social normativity*? Put in somewhat simplistic terms, the problem seems to be the following. If the agent's behavior is interpreted as instrumental action, and if it is believed that this is the fundamental way in which agents interpret each other's behavior, it seems entirely dubious why there should be anything like normative expectations.

Let's have a closer look at the matter. The problem is not that there is no place for normativity in instrumentalist accounts of practical reasons. Those accounts do imply normativity, but the kind of normativity at stake here is exclusively a matter of the relation between what an agent *wants* and what he or she actually *does*. *If* agent A wants x, he *should* (by virtue of the normativity of rationality) do y, i.e. choose the appropriate means to realize x (if there are no conflicting meta-preferences, or overriding desires). The normativity of y is *hypothetical:* it is something A should do *insofar* – and *only* insofar – as he or she wants x (or anything else that is best realized by y). Within this setting, there is no telling *what it is* that A should *want*. It is true that Weber is not Hume, for whom "reason is (. . .) the slave of passion". In contrast to Hume's passions, Weber's goals are subject to rational constraints – but only insofar as there are inconsistencies concerning the values of the agent. If consistency of the entire range of consequences with the agent's values is integrated into the concept of goal-effectiveness, Weber's famous claim that there is no rational evaluation of one's values can be transferred to one's goals (i.e. insofar as goals are value-consistent effects of the behavior that are weighed against any unintended consequences).

The decisive point, however, is yet another one. The range of purely instrumental normativity is limited to a purely monological space: instrumental rationality is a matter of the solitary relation between an agent and his or her goals. It ties the agent's behavior to his or her ends, and not one agent's behavior to another agent's. There is, it seems, nothing social about this normativity. Thus this kind of normativity seems utterly unfit to explain how rational agents should come to have normative expectations concerning each other's behavior. Instrumental agents might think of each other as rational or irrational, depending on the effectiveness of their pursuit of their goals. Insofar as there is no reason to think that other agents *should* pursue whatever goal they happen to have, there seems no reason normatively to expect them to pursue their goals *rationally*, either. It seems that the only person of whom a purely instrumentally rational agent can expect rational behavior in a *normative* way is the agent him- or herself, because rational agents cannot be indifferent as to the question of whether or not *their own goals* are achieved. Thus instrumentally rational agents cannot take a purely cognitive stance towards their own (future) behavior. When their mind is set on a goal, they have to *expect* themselves to pursue their goals effectively. It is true that agents might *take into account* the weakness of their own will, and might suspect that they will not be able to follow through with their own

plan. Thus there are cognitive components in their expectations concerning their future behavior. But their expectations cannot be *entirely* and *exclusively* cognitive. There *has* to be a normative component, at least insofar as the agent's goals are not just desired states of affairs, but *conditions of satisfaction of the agent's intentions*, i.e. whatever has to be the case for an agent to have done what he or she intended to do. In other words, to have an *intention* implies to have *some* normative expectations towards one's own behavior.

It appears, however, that within an instrumentalist account of practical reason the agent's own behavior is the *only* suitable address for normative expectations. The only genuinely normative relation an instrumentally rational agent has is the solitary relation to him- or herself.

This does not, of course, mean that instrumentally rational agents will be entirely *indifferent* to other agents' irrationalities. Where this seems useful in the light of their own plans, they will encourage other agents to be rational, too. But just as other people's values have to be taken into account, so do other people's irrationalities. There is no reason to *count on* each other's rationality, and to act on normative expectations regarding each other's rational behavior.

These considerations lead to a view of Weber's problem that is very different from either Parsons' or Habermas'. The main problem of the attempt to integrate a theory of social normativity into an instrumental account of practical reason is not the focus on action (as systems theory suggests), nor is it a matter of the focus on instrumental success as such (as Habermas believes). Rather, the problem can be narrowed down much further. It lies in the way in which, within Weber's account at least, action (in terms of meaningful behavior) is related to instrumental success (in terms of conditions of satisfaction of intentions). Just as Weber conceives of action as a matter of solitary agents, he thinks that the goals in question have to be solitary agent's goals, too. This *individualism about goals* is particularly obvious in the first introduction of the concept of instrumental rationality in the second paragraph of *Economy and Society*. Instrumentally rational action is "determined by expectations as to the behavior of objects in the environment and of other human beings; these expectations are used as 'conditions' or 'means' for the attainment of the actor's *own* rationally pursued and calculated ends" (WG 12). The emphasis is mine, for my concluding question is the following: could it be that *this* is the cause of Weber's problems: neither his focus on action, nor his alleged instrumentalism, but his *individualism about goals?*

Now a days, the place where the role of goals in action is most hotly debated is Rational Choice Theory. In this context, Amartya Sen has repeatedly questioned a more or less tacit assumption of the orthodox model of practical reason that has been accepted at least from Max Weber's days up to current Rational Choice Theory. Sen calls the assumption in question *self-goal choice*. Self-goal choice does not imply that the agent's goal is focused on his or her own self-interest. The assumption that agents aim at furthering other people's well-being is compatible with self-goal choice. The limitation at stake here is of a different nature. Self-goal choice is the claim that agents have to be interpreted as pursuing *their own* goals. But how could

this ever be otherwise? How could agents choose anything else than their own goals? How could goals be anything other than the agent's own?

Sen claims that there is a type of action, which he calls *committed action*, which is in a sense *goal oriented*, but not towards the agent's own goals. Sen is clearly aware of how strange and indeed counterintuitive his claim might appear. For even in paradigmatic cases of altruistic action, self-goal choice does not appear to be violated, because the altruists' own goal seems to be to benefit other agents, which is his own (altruistic) goal. As Sen himself puts this objection to his claim, "it might appear that if I were to pursue anything other than what I see as my own goals, then I am suffering from an illusion; these other things *are* my goals, contrary to what I might believe." (Sen 2002: 212)

Thus it is hardly surprising that most critics do not accept Sen's claim.[15] I think this defense of Rational Choice has a point against Sen, but only as far as the only alternative to the agent's individual goals are other agent's individual goals. In contrast to this, it seems to me that the merit of Sen's claim is to point our attention to another class of goals. The fact that agents cannot pursue other people's goals without making them their own does not mean that all goals which agents pursue are their own goals. Agents do not just have their own goals. They have *shared* goals, too.[16] Interestingly, it seems that the early Parsons had similar ideas. In his critique of the "atomism" of both utilitarian and contractual versions of practical reason, Parsons has made the very same point, emphasizing "that men's ends should not be separate, and either forcibly restrained or miraculously compatible, but in fact, in a given society, *held in common*." (Parsons 1934: 158; my emphasis)

It seems that the existence of shared goals opens up an alternative perspective on the relation between instrumental action and social normativity, one that makes a solution to Weber's problem of having to base a theory of social normativity on an instrumentalist theory of action possible. Where agents pursue a shared goal, instrumental normativity breaks out of the cage of the solitary relation between an individual agent and his or her own goals, and takes on a social meaning. As far as agents pursue shared goals, the instrumental rationality of other agents concerns them just as directly as their own instrumental rationality does. Just as solitary rational agents cannot be indifferent as to the achievement of their own goals, and cannot take a purely cognitive stance towards their own future behavior, they are now in normative relations to each other. If A's individual goal x_i is a *reason for him* to do y (as far as y is the best means to realize x_i), agents acting in pursuit of the *shared goal* x_s can mutually expect rational behavior (i.e. y) *from each other*, because y is the best means to realize x_s.

In other words, shared goals appear to be the point at which merely instrumental requirements turn into social normativity. Just as the relation of individual agents to themselves cannot be limited to cognitive expectations, the relations *between* agents acting in pursuit of shared goals have to include a normative element, too. They have to expect each other to choose the suitable means *normatively*. They do

[15] Cf. e.g. Pettit (2002).

[16] For a similar reconstruction of Sen's concept of committed action see Anderson (2001).

not just take each other's behavior into account; in the way the archer takes into account the wind when she aims at her target.[17] The behavior of the other agents is not just another *restriction* that has to be taken into account within one's own calculation of suitable means. And much less do agents who act in pursuit of a shared goal simply manipulate each other as means to their own goals, as Jean-Paul Sartre thinks. Rather, these agents *count on each other* (*as well as on their own future selves*), and their shared goals provide them with a point from which they can critically assess other people's behavior as well as their own.

If pre-discursive instrumental action is possible: why should this be limited to the pursuit of *individual* goals? Why should pre-discursive instrumental action not extend to the pursuit of shared goals? Especially in light of recent findings in child psychology, which has shown how important shared intentionality is even for infants, which are not yet capable of speech, and how deeply rooted the motivation for cooperation is in our psychologies (cf. Tomasello 1998), this does not seem implausible at all. If this is right, then the solution of Weber's problem is to be found in a theory of shared goals. And if the primary role goals have for action is that of conditions for the satisfaction of intentions, we need an account of *shared intentions* to solve Weber's problem. The most important amendment to be made to Weber's action theory is this. If it is a characteristic of instrumental action that it is "determined by [cognitive] expectations as to the behavior of objects in the environment and of other human beings" which are "used as 'conditions' or 'means' for the attainment of the actor's *own* rationally pursued and calculated ends" (WG 12), consensual action is, at the most basic level of the phenomenon, determined by *normative* expectations regarding the contributions of other agents for the agents' rationally pursued and calculated *shared* ends.

[17] Oddly enough, this is what Pettit seems to think; cf. Pettit (2005: 20).

Bibliography

Abrams, D. and Hogg, M. A. (eds.) 1990. "An Introduction to the Social Identity Approach." In *Social Identity Theory: Constructive and Critical Advances*. New York: Harverster, pp. 1–9.

Anderson, B. 1991. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London: Verso.

Anderson, E. 2001. "Unstrapping the Straitjacket of 'Preference': A Comment on Amartya Sen's Contributions to Philosophy and Economics." *Economics and Philosophy* 17: 21–38.

Angehrn, E. 1983. "Handlungserklärung und Rationalität: Zur Methodologie Max Webers." *Zeitschrift für philosophische Forschung* 37: 341–362.

Arrow, K. J. 1994. "Methodological Individualism and Social Knowledge." *The American Economic Review* 84/2: 1–9.

Augner, R. (ed.) 2000. *Darwinizing Culture. The Status of Memetics as a Science*. Oxford: Oxford University Press.

Bacharach, M. 1998. "Interactive Team-Reasoning: A Contribution to the Theory of Cooperation." *Research in Economics* 58: 117–147.

Bacharach, M. 2006. *Beyond Individual Choice. Teams and Frames in Game Theory*. Princeton, NJ: Princeton University Press.

Baier, A. C. 1970. "Act and Intent." *Journal of Philosophy* 67: 648–658.

Baier, A. C. 1997a. *The Commons of the Mind*. Chicago, IL: Open Court.

Baier, A. C. 1997b. "Doing Things with Others: The Mental Commons." In L. Alanen, S. Heinämaa, and T. Wallgren (eds.), *Commonality and Particularity in Ethics*. London: Macmillan, pp. 15–44.

Batson, C. D. 1994. "Prosocial Motivation: Why Do We Help Others?" In A. Tesser (ed.), *Advanced Social Psychology*. Boston, MA: McGraw-Hill, pp. 333–381.

Baurmann, M. 1996. *Der Markt der Tugend: Recht und Moral in der liberalen Gesellschaft*. Tübingen: J.C.B. Mohr.

Baxter, D. 2005. "Altruism, Grief, and Identity". *Philosophy and Phenomenological Research* LXX/2: 371–383.

Becher, E. 1917. *Die fremddienliche Zweckmäßigkeit der Pflanzengallen und die Hypothese eines überindividuellen Seelischen*. Leipzig: Veit.

Becher, E. 1921. *Geisteswissenschaften und Naturwissenschaften*. Munich: Duncker & Humblot.

Bedford, E. [1956/57] 2003. "Emotions." In R. Solomon (ed.), *What Is an Emotion? Classic and Contemporary Readings*. New York: Oxford University Press, pp. 207–216.

Belke, I. (ed.) 1971. "Einleitung." In *Moritz Lazarus, and Heymann Steinthal: Die Begründer der Völkerpsychologie in ihren Briefen*, Vol. 1. Tübingen: J. C. B. Mohr, pp. xiii–cxlii.

Benn, S. I. and Gaus, G. F. 1986. "Practical Rationality and Commitment." *American Philosophical Quarterly* 23: 255–266.

Binswanger, L. 1922. *Einführung in die Probleme der allgemeinen Psychologie*. Berlin: Springer.

Blackmore, S. 1999. *The Meme Machine*. New York/Oxford: Oxford University Press.

Bowles, S., Fehr, E. and Gintis, H. 2003. "Strong Reciprocity May Evolve With or Without Group Selection." *Theoretical Primatology Project Newsletter* 1/12.

Brandom, R. B. 1992. "Heidegger's Categories in 'Being and Time.'" In H. L. Dreyfus and H. Hall (eds.), *Heidegger: A Critical Reader*. Cambridge, MA: Blackwell, pp. 45–64.
Brandom, R. B. 1994. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.
Brandom, R. B. 1997. "Replies." *Philosophy and Phenomenological Research* LVII/1: 191–207.
Brandom, R. B. 2000. *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.
Bratman, M. E. 1987. *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.
Bratman, M. E. 1999. *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge, MA: Harvard University Press.
Bratman, M. E. 2007. *Structures of Agency*. Oxford: Oxford University Press.
Brodie, R. 1996. *Viruses of the Mind: The New Science of the Meme*. Seattle, WA: Integral Press.
Broome, J. 2000. "Normative Requirements." In J. Dancy (ed.), *Normativity*. Oxford: Blackwell, pp. 78–99.
Broome, J. 2001. "Are Intentions Reasons? And How Should We Cope with Incommensurable Values?" In C. Morris and A. Ripstein (eds.), *Practical Rationality and Preference: Essays for David Gauthier*. Cambridge: Cambridge University Press, pp. 98–120.
Bühler, K. [1927]. 2000. *Die Krise der Psychologie*. Weilerwist: Velbrück.
Camerer, C. F. and Fehr, E. 2004. "Measuring Social Norms and Preferences. Using Experimental Games: A Guide for Social Scientists." In J. Henrich et al. (eds.), *Foundation of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press, pp. 55–95.
Canto i Mila, N. 2002. "The Legacy of an Extinguished Discipline: On the Intellectual Relationship Between Moritz Lazarus and Georg Simmel and the Influence of the Psychology of Nations on the Constitution of Sociology." *Simmel Studies* 12: 263–280.
Carman, T. 1994. "On Being Social: A Reply to Olafson." *Inquiry* 37: 203–23.
Carman, T. 2000. "Must We Be Inauthentic?" In M. Wrathall and J. Malpas (eds.), *Heidegger, Authenticity, and Modernity: Essays in Honor of Hubert L. Dreyfus*, Vol. 1. Cambridge, MA: Harvard University Press, pp. 13–28.
Celano, B. 1999. "Collective Intentionality, Self-Referentiality, and False Beliefs: Some Issues Concerning Institutional Facts." *Analyse und Kritik* 21: 237–250.
Clore, G. L. 1994. "Why Emotions are Never Unconscious." In P. Ekman and R. J. Davidson (eds.), *The Nature of Emotion. Fundamental Questions*. Oxford: Oxford University Press, pp. 285–290.
Cohn, J. 1919. *Geist der Erziehung: Pädagogik auf philosophischer Grundlage*. Leipzig/Berlin: Teubner.
Coleman, J. S. 1990. *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.
Coleman, J. S. 1992. "Social Institutions and Social Theory." In P. Hamilton (ed.), *Talcott Parsons: Critical Assessments*, Vol. 2. London: Routledge, pp. 43–51.
Collingwood, R. G. [1942] 1947. *The New Leviathan*. Oxford: Clarendon.
Cooley, C. H. [1902/1905] 1956. *Social Organization: Human Nature and the Social Order*. Glencoe, IL: Free Press.
Darwin, C. R. [1859] 1975. *The Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*. New York: AMC Press.
Darwin, C. R. 1872. *The Expression of the Emotions in Man and Animals*. London: John Murray.
Davidson, D. 1963. "Actions, Reasons, and Causes." *The Journal of Philosophy* LX (23): 685–700.
Davis, J. B. 2003. *The Theory of the Individual in Economics*. London: Routledge.
Dawkins, R. 1976. *The Selfish Gene*. New York/Oxford: Oxford University Press.
Dennett, D. C. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology*. Brighton: Harvester Press.
Dennett, D. C. 1991. *Consciousness Explained*. Boston, MA: Little Brown.
Dennett, D. C. 1993. *Consciousness Explained*. London: Penguin Books.
Dennett, D. C. 1995. *Darwin's Dangerous Idea*. London: Penguin Books.

Descartes, R. [1641] 1915. *Meditationes de prima philosophia*. Leipzig: Felix Meiner.
Dilthey, W. 1923. *Einleitung in die Geisteswissenschaften: Versuch einer Grundlegung für das Studium der Gesellschaft und der Geschichte. Wilhelm Dilthey gesammelte Schriften*, Vol. I. Leipzig/Berlin: Teubner.
Dreyfus, H. L. 1991. Being-in-the-World: A Commentary on Heidegger's Being and Time, Division 1. Cambridge, MA: MIT Press.
Dreyfus, H. L. 1993. "Heidegger's Critique of the Husserl/Searle Account of Intentionality." *Social Research* 60: 17–38.
Dreyfus, H. L. 1995. "Interpreting Heidegger on Das Man." *Inquiry* 38: 423–430.
Dreyfus, H. L. 1999. "Coping with Things-in-Themselves: A Practice-Based Phenomenological Argument for Realism." *Inquiry* 42: 49–78.
Dreyfus, H. L. 2000. "Could Anything Be More Intelligible than Everyday Intelligibility? Reinterpreting Division I of Being and Time in the Light of Division II." In J. E. Faulconer and M. A. Wrathall (eds.), *Appropriating Heidegger*. Cambridge, MA: Harvard University Press, pp. 155–174.
Dupré, J. 2003. Review of Augner (ed.) 2000. *Mind and Language* 18: 220–224.
Durkheim, E. [1898] 1994. "On Social Facts." In M. Martin and L. McIntyre (eds.), *Readings in the Philosophy of Social Science*. Cambridge, MA: MIT Press, pp. 433–440.
Elster, J. 1979. *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.
Elster, J. 1985. "The Nature and Scope of Rational-Choice Explanation". In E. LePore and B. McLaughlin (eds.), *Actions and Events: Perspectives on Donald Davidson*. Oxford: Blackwell, pp. 60–72.
Elster, J. 1989. *The Cement of Society: A Study of Social Order*. Cambridge: Cambridge University Press.
Fehr, E. and Fischbacher, U. 2003. "The Nature of Human Altruism." *Nature* 425: 785–791.
Fehr, E. and Fischbacher, U. 2004. "Third-Party Punishment and Social Norms." *Evolution of Human Behavior* 25: 63–87.
Fehr, E. and Gächter, S. 2002. "Altruistic Punishment in Humans." *Nature* 415: 137–140.
Flood, M. M. 1958. "Some Experimental Games." *Management Science* 5: 5–26.
Freud, S. [1916/17] 1957. *Introductory Lectures on Psychoanalysis*. Sigmund Freud Standard Edition Vol. 15. New York: W.W. Norton.
Garfinkel, A. 1981. *Forms of Explanation. Rethinking the Questions in Social Theory*. New Haven, CT: Yale University Press.
Gatherer, D. 1998. "Why the 'Thought Contagion' Metaphor is Retarding the Progress of Memetics." *Journal of Memetics* 2.
Gauthier, D. 1975. "Coordination." *Dialogue* 14: 195–221.
Geiger, T. 1987. *Vorstudien zu einer Soziologie des Rechts*. Berlin: Duncker und Humblot.
Gierke, Otto von. 1915. *Der deutsche Volksgeist im Kriege*. Stuttgart/Berlin: Deutsche Verlags-Anstalt.
Gilbert, M. 1989 1992. (reprint). *On Social Facts*. Princeton, NJ: Princeton University Press.
Gilbert, M. 1996. *Living Together: Rationality, Sociality, and Obligation*. Lanham MD: Rowman & Littlefield.
Gilbert, M. 1997. "Group Wrongs and Guilt Feelings." *The Journal of Ethics* 1: 65–84.
Gilbert, M. 2000. *Sociality and Responsibility: New Essays in Plural Subject Theory*. Lanham MD: Rowman & Littlefield.
Gilbert, M. 2002a. "Acting together." In G. Meggle (ed.), *Social Facts and Collective Intentionality*. German Library of Sciences, Philosophical Research, Vol. 1. Frankfurt am Main: Hänsel-Hohenhausen, pp. 53–71.
Gilbert, M. 2002b. "Collective Guilt and Collective Guilt Feelings." *The Journal of Ethics* 6: 115–143.
Gold, N. (ed.) 2005. *Teamwork: Multi-Disciplinary Perspectives*. New York: Macmillan.
Gold, N. and Sugden, R. 2007. "Theories of Team-Agency." In F. Peter and H. B. Schmid (eds.), *Rationality and Commitment*. Oxford: Oxford University Press, pp. 280–311.

Goldie, P. 2000. *The Emotions: A Philosophical Exploration*. Oxford: Oxford University Press.
Goyal, S. and Janssen, M. C. 1996. "Can We Rationally Learn to Coordinate?" *Theory and Decision* 40: 29–40.
Graber, E. 1981. "Translator's Introduction to Max Weber's Essay on Some Categories of Interpretive Sociology." *Sociological Quarterly* 22/2: 145–150.
Grice, H. P. 1971. *Intention and Uncertainty*. Oxford: Oxford University Press.
Grossmann, A. 2000. "Volksgeist – Grund einer praktischen Welt oder metaphysische Spukgestalt? Anmerkungen zur Problemgeschichte eines nicht nur Hegelschen Theorems." In A. Grossmann and C. Jamme (eds.), *Metaphysik in der praktischen Welt: Perspektiven im Anschluß an Hegel und Heidegger*. Amsterdam: Rodopi, pp. 60–77.
Grossmann, A. 2001. Volksgeist; Volksseele. In *Historisches Wörterbuch der Philosophie*, Vol. 11. Basel: Schwabe, pp. 1102–1107.
Gumplowicz, L. 1928. *Soziologische Essays*. Wien: Universitäts-Verlag Wagner.
Habermas, J. (ed.) [1971/72] 1984. "Vorlesungen zu einer sprachtheoretischen Grundlegung der Soziologie." In *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns*. Frankfurt a. M.: Suhrkamp.
Habermas, J. 1981. *Theorie des kommunikativen Handelns. Vol. 1: Handlungsrationalität und gesellschaftliche Rationalisierung*. Frankfurt a. M.: Suhrkamp.
Habermas, J. 1987. *Theory of Communicative Action*, Vol. I and II, translated by T. McCarthy. Boston, MA: Beacon press.
Habermas, J. 1992. *Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und demokratischen Rechtsstaats*. Frankfurt a. M.: Suhrkamp.
Habermas, J. 1998. *Die postnationale Konstellation: Politische Essays*. Frankfurt a. M.: Suhrkamp.
Habermas, J. (ed.) 2001. "Reflections on the Linguistic Foundation of Sociology. The Christian Gauss Lecture, February-March". In *On the Pragmatics of Social Interaction. Preliminary Studies in the Theory of Communicative Action*. Cambridge MA: MIT Press, pp. 1–104. Originally published in German as "Vorlesungen zu einer sprachtheoretischen Grundlegung der Soziologie" In *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns*. Frankfurt am Main [1971/72] 1984.
Hamer, D. 2004. *The God Gene. How Faith Is Hardwired into Our Genes*. New York: Doubleday.
Harman, G. 1986. *Change in View*. Cambridge MA: MIT Press.
Harsanyi, J. C. and Selten, R. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge MA: MIT Press.
Hartmann, E. 1871. "Ueber das Wesen des Gesammtgeistes: Eine kritische Betrachtung des Grundbegriffes der Völkerpsychologie." *Zeitschrift für Philosophie und philosophische Kritik* Neue Folge 58: 16–32.
Hartshorne, C. 1942. "Elements of Truth in the Group-Mind Concept." *Social Research* 9: 248–265.
Haugeland, J. 1982. "Heidegger on Being a Person." *Nous* XVI: 15–26.
Haugeland, J. 1992. "Dasein's Disclosedness." In H. L. Dreyfus and H. Hall (eds.), *Heidegger: A Critical Reader*. Cambridge, MA: Blackwell, pp. 27–44.
Haugeland, J. 2000. "Truth and Finitude." In M. Wrathall and J. Malpas (eds.), *Heidegger, Authenticity, and Modernity: Essays in Honor of Hubert L. Dreyfus*, Vol. 1. Cambridge, MA: Harvard University Press, pp. 43–77.
Heath, J. 2005. "Methodological Individualism." *Stanford Encyclopedia of Philosophy*. < http://www.plato.stanford.edu >.
Heidegger, M. [1925] 1979. *Prolegomena zur Geschichte des Zeitbegriffes, Gesamtausgabe, Vol. 20*. Frankfurt a.M.: Klostermann.
Heidegger, M. [1927] 1996. *Being and Time. A Translation of* Sein und Zeit. Translated by J. Stambaugh. Albany, NY: State University of New York Press.
Heidegger, M. [1928/29] 1996. *Einleitung in die Philosophie. Gesamtausgabe, Vol. 27*. Frankfurt a.M.: Klostermann.
Heidegger, M. [1933] 1983. *Die Selbstbehauptung der deutschen Universität*. Frankfurt a.M.: Klostermann.

Heidegger, M. [1934] 1998. *Logik als die Frage nach dem Wesen der Sprache, Gesamtausgabe Vol. 38*. Frankfurt a.M.: Klostermann.
Heidegger, M. [1936–38] 1989. *Beiträge zur Philosophie (vom Ereignis), Gesamtausgabe, Vol. 65*. Frankfurt a.M.: Klostermann.
Heidegger, M. [1938/39] 1997. *Besinnung. Gesamtausgabe, Vol. 66*. Frankfurt a.M.: Klostermann.
Helm, B. 2008. "Plural Agents." *Noûs* 42: 17–49.
Henrich, J. et al. (eds.) 2004. *Foundation of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press.
Heyer, K. [1961] 1990. *Wer ist der deutsche Volksgeist?* Basel: Perseus Verlag.
Hindriks, F. 2002. "Social Ontology, Collective Intentionality, and Ockhamian Skepticism." In G. Meggle (ed.), *Social Facts and Collective Intentionality*. German Library of Sciences, Philosophical Research, Vol. 1. Frankfurt am Main: Hänsel-Hohenhausen, pp. 125–149.
Hirschman, A. O. 1970. *Exit, Voice, and Loyalty*. Cambridge, MA: Harvard University Press.
Hochschild, A. R. 1979. "Emotion Work, Feeling Rules and Social Structure." *American Journal of Sociology* 85: 551–575.
Hofstadter, D. C. 1985. *Metamagical Themas*. New York: Basic Books.
Hollis, M. (ed.) 1987. "External and Internal Reasons." In *The Cunning of Reason*. Cambridge: Cambridge University Press, pp. 74–94.
Hollis, M. and Sugden, R. 1993. "Rationality in Action." *Mind* 102: 1–35.
van Hooft, S. 1994. "Scheler on Sharing Emotions." *Philosophy Today* 38/1: 18–28.
Hornsby, J. 1997. "Collectives and Intentionality." *Philosophy and Phenomenological Research* 57: 429–434.
Hume, D. (1739/40) 2000. *A Treatise of Human Nature*. D. F. Norton and M. J. Norton (eds), Oxford: Oxford University Press.
Hurley, S. L. 1989. *Natural Reasons: Personality and Polity*. Oxford: Oxford University Press.
Jahoda, G. 2002. "The Ghosts in the Meme Machine." *History of the Human Sciences* 15: 55–68.
Janssen, M. C. W. 2000. "Towards a Justification of the Principle of Coordination." *Tinbergen Institute Discussion Papers* No. 00–017/1.
Janssen, M. C. W. 2001a. "On the Principle of Coordination." *Economics and Philosophy* 17: 221–234.
Janssen, M. C. W. 2001b. "Rationalizing Focal Points." *Theory and Decision* 50: 119–148.
Jellinek, G. 1914. *Allgemeine Staatslehre*. Berlin: O. Häring.
Johansson, I. 2003. "Searle's Monadological Construction of Social Reality." *American Journal of Economics and Sociology* 62: 233–255.
Kagel, J. K. and Roth, A. E. (eds.) 1995. *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.
Kenny, A. 1976. Human Abilities and Dynamic Modalities. In J. Manninen and R. Tuomela (eds.), *Essays on Explanation and Understanding. Studies in the Foundations of Humanities and Social Sciences*. Dordrecht: D. Reidel, pp. 209–232.
Kitcher, P. 1998. "Psychological Altruism, Evolutionary Origins, and Moral Rules." *Philosophical Studies* 89: 283–316.
Knobe, J. and Prinz, J. 2007. "Intuitions About Consciousness." *Phenomenology and Cognitive Science* 7: 67–85.
Köhnke, K.-C. 1996. *Der junge Simmel in Theoriebeziehungen und sozialen Bewegungen*. Frankfurt a. M.: Suhrkamp.
Kollock, P. 1998. "Transforming Social Dilemmas: Group Identity and Co-operation." In P. A. Danielson (ed.), *Modeling Rationality, Morality, and Evolution*. New York: Oxford University Press, pp. 185–209.
Kronfeld, A. 1920. *Das Wesen der psychiatrischen Erkenntnis: Beiträge zur allgemeinen Psychiatrie*. Berlin: Springer.
Kutz, C. 2000a. "Acting Together." *Philosophy and Phenomenological* Research 61: 1–31.
Kutz, C. 2000b. *Complicity: Ethics and Law for a Collective Age*. Cambridge: Cambridge University Press.

Larenz, K. 1935. "Volksgeist und Recht: Zur Revision der Rechtsanschauung der Historischen Schule." *Zeitschrift für Deutsche Kulturphilosophie* 1: 39–60.

Latour, B. 2002. "Gabriel Tarde and the End of the Social." In P. Joyce (ed.), *The Social in Question. New Bearings in History and the Social Sciences*. London: Routledge, pp. 117–32.

Lauer, D. 2005. "Normativität, Inferentialität und Verstehen." In G. W. Bertram et al. (eds.), *Intersubjektivité et pratique: Contributions à l'étude des pragmatismes dans la philosophie contemporaine*. Paris: L'Harmattan, pp. 75–89.

Lazarus, M. 1850. *Die sittliche Berechtigung Preußens in Deutschland*. Berlin.

Lazarus, M. 1880. *Was heißt national? Ein Vortrag*. Berlin: Philo.

Lazarus, M. [1851–65] 2003. *Gundzüge der Völkerpsychologie und Kulturwissenschaft*. K.-C. Köhnke (ed.), Hamburg: Meiner.

Lazarus, M. and Steinthal, H. 1860. "Einleitende Gedanken über Völkerpsychologie als Einladung zu einer Zeitschrift für Völkerpsychologie und Sprachwissenschaft." *Zeitschrift für Völkerpsychologie und Sprachwissenschaft* 1: 1–73.

Lazarus, M. and Steinthal, H. 1986. *Die Begründer der Völkerpsychologie in ihren Briefen*, vol. II/2. Mit einer Einleitung herausgegeben von Ingrid Belke (ed.), Tübingen: J.C.B. Mohr.

Lazarus, N. 1910. *Ein deutscher Professor in der Schweiz*. Berlin: Ferd. Dümmler.

LeDoux, J. E. 1994. "Emotional Processing, but not Emotions, can Occur Unconsciously." In P. Ekman and R. J. Davidson (eds.), *The Nature of Emotion: Fundamental Questions*. Oxford: Oxford University Press, pp. 291–292.

Levesque, H. J. and Cohen, P. R. 1991. "Teamwork." *Noûs* 25: 487–512.

Lévinas, E. 1991. *Entre nous. Essais sur le penser-à-l'autre*. Paris: Bernard Grasset.

Lewis, D. K. 1969. *Convention: A Philosophical Study*. Oxford: Blackwell.

Leys, R. 1993. "Mead's Voices: Imitation as Foundation, or, the Struggle Against Mimesis." *Critical Inquiry* 19: 277–307.

Lipps, T. 1903. *Grundlegung der Ästhetik*. Hamburg/Leipzig: Voß.

Löwenstein, K. K. 1911. "Über den Akt des Könnens"und seine Bedeutung für Praktik, Didaktik und Pädagogik." *Zeitschrift für wissenschaftliche pädagogische Psychologie und experimentelle Pädagogik* XII, S. 403–420.

Ludendorff, M. 1933. *Die Volksseele und ihre Machtgestalter*. München: Ludendorffs Volkswarte.

Ludwig, K. 1992. "Impossible Doings". *Philosophical Studies* 65: 257–281.

Luhmann, N. 1984. *Soziale Systeme: Grundriß einer allgemeinen Theorie*. Frankfurt a. M.: Suhrkamp.

Luhmann, N. 1987. *Beiträge zur funktionalen Differenzierung der Gesellschaft*. Soziologische Aufklärung Vol. 4. Opladen: Westdeutscher Verlag.

Lynch, A. 1996. *Thought Contagion: How Belief Spreads Through Society*. New York: Basic Books.

Maffesoli, M. 1978. "Dynamique de la dissidence." *Cahiers Internationaux de Sociologie* 64: 103–111.

Marcel, A. 2003. "The Sense of Agency. Awareness and Ownership of Action." In J. Roessler and N. Eilan (eds.), *Agency and Self-Awareness*. Oxford: Oxford University Press, pp. 48–93.

Marsden, P. 1998. "Memetics and Social Contagion: Two Sides of the Same Coin?" *The Journal of Memetics* 2: http://cfpm.org/jom-emit/.

Marsden, P. 2000. "Forefathers of Memetics: Gabriel Tarde and the Laws of Imitation." *The Journal of Memetics* 4: http://cfpm.org/jom-emit/.

Mathiesen, K. 2002. "Searle, Collective Intentions, and Individualism." In G. Meggle (ed.), *Social Facts and Collective Intentionality. German Library of Sciences, Philosophical Research, Vol. 1*. Frankfurt a. M.: Hänsel-Hohenhausen, pp. 185–204.

Mathiesen, K. 2003. "On Collective Identity." *Protosociology* 18: 66–68.

McNaughton, D. and Rawling, P. 2004. "Duty, Rationality, and Practical Reasons." In A. R. Mele and P. Rawling (eds.), *The Oxford Handbook of Rationality*. Oxford: Oxford University Press, pp. 110–31.

Meijers, A. W. M. 1994. *Speech Acts, Communication, and Collective Intentionality: Beyond Searle's Individualism*. Utrecht: A.W.M. Meijers.

Meijers, A. W. M. 2002. "Dialogue, Understanding and Collective Intentionality." In G. Meggle (ed.), *Social Facts and Collective Intentionality*. German Library of Sciences, Philosophical Research, Vol. 1. Frankfurt a. M.: Hänsel-Hohenhausen, pp. 225–254.
Meijers, A. W. M. 2003. "Can Collective Intentionality Be Individualized?" *American Journal of Economics and Sociology* 62: 167–183.
Mele, A. F. 2003. *Motivation and Agency*. Oxford: Oxford University Press.
Milgram, S. 1974. *Obedience to Authority: An Experimental View*. New York: Harper & Row.
Miller, S. 1991. "Coordination, Salience, and Rationality." *Southern Journal of Philosophy* 29: 359–371.
Millikan, R. G. 2003. "Vom angeblichen Siegeszug der Gene und der Meme." In A. Becker et al. (eds.), *Gene, Meme und Gehirne: Geist und Gesellschaft als Natur*. Frankfurt a. M.: Suhrkamp, pp. 90–111.
Mucchielli, L. 2000. "Tardomania? Réflexions sur les usages contemporains de Tarde." *Revue d'histoire des sciences humaines* 3: 161–184.
Nagel, T. 1970. *The Possibility of Altruism*. Oxford: Clarendon Press.
Nietzsche, F. [1881] 1977. Morgenröthe. In G. Colli and M. Montinari (eds.), *Kritische Studienausgabe*, Vol. III. Berlin: Walter de Gruyter.
Nietzsche, F. 1978. *Nachgelassene Fragmente Sommer 1872 bis Ende 1874*. In G. Colli and M. Montinari (eds.), Werke. Kritische Gesamtausgabe, Vol. 3/IV. Dritte Abteilung, vierter Band. Berlin: Walter de Gruyter.
Norkus, Z. 2001. *Max Weber und Rational Choice*. Marburg: Metropolis.
Nozick, R. 2001. *Invariances: The Structure of the Objective World*. Cambridge, MA: Belknap.
Okrent, M. 1988. *Heidegger's Pragmatism*. Ithaca, NY: Cambridge University Press.
Ortega y Gasset, J. 1957. *Man and People*. New York: Norton.
Paprzycka, K. 1998. "Collectivism on the Horizon." *Australasian Journal of Philosophy* 76: 165–181.
Paprzycka, K. 2002. "The False Consciousness of Intentional Psychology." *Philosophical Psychology* 15/3: 271–295.
Parsons, T. [1937] 1949. *The Structure of Social Action: A Study in Social Theory with special Reference to a Group of Recent European Writers*. Glencoe, IL: Free Press.
Parsons, T. 1951. *The Social System*. Glencoe, IL: Free Press.
Parsons, T. (ed.) 1954. "The Prospects of Sociological Theory." In *Essays in Sociological Theory*. Revised Edition. Glencoe, IL: Free Press, pp. 348–369.
Parsons, T. 1968. "Social Interaction." In *International Encyclopedia of the Social Sciences*, Vol. 7. New York: Macmillan Reference, pp. 429–441.
Parsons, T. 1991. *The Early Essays*. C. Camic (ed.). Chicago/London: University of Chicago Press.
Parsons, T. and Shils, E. A. (eds.) 1951. "Categories of the Orientation and Organization of Action." In *Toward a General Theory of Action*. Cambridge, MA: Harvard University Press, pp. 53–109.
Parsons, T. and Shils, E. A. (eds.) 1959. "Some Fundamental Categories of the Theory of Action: A General Statement." In *Toward a General Theory of Action*. Cambridge, MA: Blackwell, pp. 3–29.
Peikoff, A. 2003. "Rational Action Entails Rational Desire: A Critical Review of Searle's Rationality in Action." *Philosophical Explorations* 4: 124–138.
Peter, F. and Schmid, H. B. (eds.) 2007. *Rationality and Commitment*. Oxford: Oxford University Press.
Pettit, P. 1996. *The Common Mind: An Essay on Psychology, Society, and Politics*. New York: Oxford University Press.
Pettit, P. 2002. "Collective Persons and Powers." *Legal Theory* 8: 443–470.
Pettit, P. 2003. "Groups with Minds of their Own." In F. Schmitt (ed.), *Socializing Metaphysics*. New York: Rowman & Littlefield, pp. 167–195.
Pettit, P. 2005. "Construing Sen on Commitment." *In Economics and Philosophy* 21/1: 15–32.
Pettit, P. and Schweikard, D. 2005. "Joint Actions and Group Agents." *Philosophy of the Social Sciences* 36/1: 18–39.

Pettit, P. and Smith, M. 2004. "Backgrounding Desires." In F. Jackson, P. Pettit, M. Smith (eds.), *Mind, Morality, and Explanation*. Oxford: Oxford University Press, pp. 269–293.
Popper, K.-R. 1962. *The Open Society and Its Enemies*. Princeton, NJ: Princeton University Press.
Poundstone, W. 1993. *Prisoner's Dilemma*. New York: Anchor Books.
Provis, C. 1977. "Gauthier on Coordination." *Dialogue* 16: 507–9.
Putnam, H. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.
Raiffa, H. 1992. "Game Theory at the University of Michigan 1948–1952." In E. R. Weintraub (ed.), *Toward a History of Game Theory*. Durham, NC: Duke University Press, pp. 165–175.
Raines, H. 1977. *My Soul Is Rested: Movement Days in the Deep South Remembered*. New York: Putnam.
Ratcliffe, M. 2005. "The Feeling of Being." *Journal of Consciousness Studies* 12/8–10: 43–60.
Reddy, R. 1980. "Individual Philanthropy and Giving Behavior." In D. Smith and J. Macaulay (eds.), *Participation in Social and Political Activities*. San Francisco, CA: Jossey-Bass, pp. 370–399.
Roffenstein, G. 1926. *Das Problem des psychologischen Verstehens*. Stuttgart: Püttmann.
Rosenberg, J. 1980. *One World and Our Knowledge of It: The Problematic of Realism in Post-Kantian Perspective*. Dordrecht: D. Reidel.
Rota, G.-C. 1989. "'Fundierung' as a Logical Concept." *The Monist* 72: 70–77.
Rotenstreich, N. 1973. "Volksgeist." In *Dictionary of the History of Ideas*, Vol. 4. New York: Scribner's Sons, pp. 490–496.
Roth, A. 2006. "Shared Agency and Contralateral Commitments." *The Philosophical Review* 113/3: 359–410.
Rothacker, E. 1920. *Einleitung in die Geisteswissenschaften*. Tübingen: J.C.B. Mohr.
Rovane, C. 1998. *The Bounds of Agency: An Essay in Revisionary Metaphysics*. Princeton, NJ: Princeton University Press.
Ryle, G. 1949. *The Concept of Mind*. London: Hutchinson.
Saaristo, A. J. 2005. Social Ontology and Agency: *Methodological Holism Naturalised*. Unpublished manuscript (doctoral dissertation to be presented to the University of London).
Sandel, M. J. 1982. *Liberalism and the Limits of Justice*. Cambridge: Cambridge University Press.
Sartre, J.-P. [1943] 1991. *L'être et le néant. Essai d'ontologie phénoménologique*. Paris: Gallimard.
Scheler, M. [1912] 1954. *The Nature of Sympathy*. Translated by P. Heath. London: Routledge and Kegan.
Scheler, M. [1912–1916] 1973. *Formalism in Ethics and Non-Formal Ethics of Values: A New Attempt toward the Foundation of an Ethical Personalism*. Evanston, IL: Northwestern University Press.
Schelling, T. C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
Schmid, H. B. 2000. *Subjekt, System, Diskurs: Edmund Husserls Begriff transzendentaler Subjektivität in sozialtheoretischen Bezügen*. Phaenomenologica Vol. 158. Dordrecht: Kluwer.
Schmid, H. B. 2003a. "Rationality-in-Relations." *American Journal of Economics and Sociology* 62: 67–101. Reprinted in L. S. Moss and D. Koepsell (eds.), *John Searle's Ideas about Social Reality: Extensions, Criticisms, and Reconstructions*. Oxford: Basil Blackwell.
Schmid, H. B. 2003b. "Can Brains in Vats Think as a Team?" *Philosophical Explorations* 6: 201–218. See Chapter 2 in this volume.
Schmid, H. B. 2003c. "Heidegger and Gilbert Ryle." In D. Thomä (ed.), *Heidegger-Handbuch*. Stuttgart: Metzler.
Schmid, H. B. 2005a. "Beyond Self Goal Choice: Amartya Sen's Analysis of the Structure of Commitment and the Role of Shared Desires." *Economics and Philosophy* 21: 51–63. See Chapter 7 in this volume.
Schmid, H. B. 2005b. "Nostrism – Social Identities in Experimental Games." *Analyse und Kritik* 27: 172–87. See Chapter 5 in this volume.
Schmid, H. B. 2005c. *Wir-Intentionalität: Kritik des ontologischen Individualismus und Rekonstruktion der Gemeinschaft*. Freiburg: Alber.
Schmid, H. B. 2005d. "Wir-identität: Reflexiv und vorreflexiv." *Deutsche Zeitschrift für Philosophie* 53/3: 365–376.

Schmid, H. B. 2007. "Rationalizing Coordination." In M. D. White and B. Montero (eds.), *Economics and the Mind*. London: Routledge, pp. 159–179. See Chapter 6 in this volume.
Schneider, C. M. 1990. *Wilhelm Wundts Völkerpsychologie*. Bonn: Bouvier.
Schopenhauer, A. [1841] 1988. Preisschrift über die Grundlage der Moral. Arthur Schopenhauer Kleinere Schriften. Zürich: Haffmanns, pp. 552–606.
Schueler, G. F. 1995. *Desire: Its Role in Practical Reason and the Explanation of Action*. Cambridge, MA: MIT Press.
Schumpeter, J. A. 1908. *Das Wesen und der Hauptinhalt der theoretischen Nationalökonomie*. Leipzig: Dunker und Humblot.
Schumpeter, J. A. 1909. "On the Concept of Social Value." *The Quarterly Journal of Economics* 2: 213–232.
Schütz, A. [1931] 1991. *Der sinnhafte Aufbau der sozialen Welt: Eine Einleitung in die verstehende Soziologie*. Frankfurt a. M.: Suhrkamp.
Searle, J. R. 1983. *Intentionality. An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.
Searle, J. R. 1990. "Collective Intentions and Actions." In P. Cohen, M. Morgan, and M. Pollack (eds.), *Intentions in Communication*. Cambridge, MA: MIT Press, pp. 401–415.
Searle, J. R. 1995. *The Construction of Social Reality*. New York: Free Press.
Searle, J. R. 1997a. "Replies to Critics." *Philosophy and Phenomenological Research* 57: 449–451.
Searle, J. R. 1997b. "Replies to Critics of the Construction of Social Reality." *History of the Human Sciences* 10: 103–110.
Searle, J. R. 1998a. *Mind, Language and Society: Philosophy in the Real World*. New York: Basic Books.
Searle, J. R. 1998b. "Social Ontology and the Philosophy of Society." *Analyse und Kritik* 20: 143–158.
Searle, J. R. 2001a. *Rationality in Action*. Cambridge, MA: MIT Press.
Searle, J. R. 2001b. "Meaning, Mind and Reality." *Revue internationale de philosophie* 55: 173–179.
Segal, G. M. A. 2000. *A Slim Book about Narrow Content*. Cambridge, MA: MIT Press.
Sellars, W. 1974. *Essays in Philosophy and Its History*. Dordrecht: D. Reidl.
Sellars, W. 1980. "On Reasoning About Values." *American Philosophical Quarterly* 17: 81–101.
Sellars, W. 1992. Science and Metaphysics. Variations on Kantian Themes. Atascadero, CA: Ridgeview.
Sen, A. K. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy and Public Affairs* 6/4: 317–344.
Sen, A. K. 1985. "Goals, Commitment, and Identity." *Journal of Law, Economics, and Organization* 1/2: 341–355. Reprinted 2002 in A. K. Sen, *Rationality and Freedom*. Cambridge, MA: Harvard University Press, pp. 206–225.
Sen, A. K. 1987. *On Ethics and Economics*. Oxford: Blackwell.
Sen, A. K. 1995. "Is the Idea of Purely Internal Consistency of Choice Bizarre?" In J. E. J. Altham and J. Harrison (eds.), *World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams*. Cambridge: Cambridge University Press, pp. 19–31.
Sen, A. K. 1999. *Reason Before Identity. The Romanes Lecture for 1998*. Oxford: Oxford University Press.
Sen, A. K. 2002. *Rationality and Freedom*. Cambridge, MA: Harvard University Press.
Sen, A. K. 2004. "*Social Identity." Revue de philosophie économique* 9: 7–27.
Simmel, G. [1908] 1983. *Soziologie: Untersuchungen über die Formen der Vergesellschaftung*. Berlin: Duncker und Humbolt.
Smith, A. [1759] 2000. *The Theory of Moral Sentiments*. Amherst, Prometheus Books.
Sober, E. and Wilson, D. S. 1998. *Unto Others: The Evolution and Psychology of Unselvish Behavior*. Cambridge, MA: Harvard University Press.
Solomon, R. [1975] 2003. "Emotions and Choice." In *Not Passion's Slave: Emotions and Choice*. Cambridge: Cambridge University Press, pp. 3–24.

Solomon, R. 1993. "The Philosophy of Emotions." In M. Lewis and J. M. Haviland (eds.), *Handbook of Emotions*. New York: Guilford, pp. 3–15.
Solomon, R. 2006. *The Passions: Philosophy and the Intelligence of Emotions*. Chantilly, VA: The Teaching Company.
Spann, O. 1921. *Der wahre Staat: Vorlesungen über Abbruch und Neubau der Gesellschaft*. Leipzig: Quelle und Meyer.
Sparks, H. 1997. "Dissident Citizenship: Democratic Theory, Political Courage, and Activist Women." *Hypatia* 12/4: 74–110.
Spence, S. A. 2001. "Alien Control: From Phenomenology to Cognitive Neurobiology". *Philosophy, Psychiatry, & Psychology* 8(2/3): 163–172.
Stein, E. 1917. *Zum Problem der Einfühlung*. Halle: Niemeyer.
Stein, E. 1922. "Beiträge zur philosophischen Begründung der Psychologie und der Geisteswissenschaften." In E. Husserl (ed.), *Jahrbuch für Philosophie und Phänomenologische Forschung, Vol. V*. Halle: Niemeyer, pp. 1–283.
Stocker, M. 2003. "The Irreducibility of Affectivity." In R. Solomon (ed.), *What Is an Emotion? Classic and Contemporary Readings*. New York: Oxford University Press, pp. 258–264.
Stoutland, F. 1997. "Why Are Philosophers of Action so Anti-Social?" In L. Alanen, S. Heinämaa, and T. Wallgren (eds.), *Commonality and Particularity in Ethics*. London: Macmillan, pp. 45–74.
Stoutland, F. 2002. "Review of Bratman: Faces of Intention." *Philosophy and Phenomenological Research* 65: 238–241.
Strawson, P. F. 1992. *Analysis and Metaphysics*. Oxford: Oxford University Press.
Strub, C. 2005. *Sanktionen des Selbst: Zur normativen Praxis sozialer Gruppen*. Freiburg/München: Alber.
Stueber, K. 2006. *Rediscovering Empathy. Agency, Folk Psychology, and the Human Sciences*. Cambridge, MA: MIT Press.
Sugden, R. 1993. "Thinking as a Team: Towards an Explanation of Nonselfish Behavior." *Social Philosophy and Policy* 10: 69–89.
Sugden, R. 1995. "A Theory of Focal Points." *Economic Journal* 105: 533–50.
Sugden, R. 1996. "Rational Coordination." In F. Farina, F. Hahn, and S. Vanucci (eds.), *Ethics, Rationality and Economic Behavior*. Oxford: Clarendon Press, pp. 244–262.
Sugden, R. 2000. "Team Preferences." *Economics and Philosophy* 16: 175–204.
Sugden, R. 2002. "Beyond Sympathy and Empathy: Adam Smith's Concept of Fellow-Feeling." *Economics and Philosophy* 18: 63–87.
Sugden, R. 2003. "The Logic of Team Reasoning." *Philosophical Explorations* 6: 165–181.
Sunstein, C. R. 2003. *Why Societies Need Dissent*. Cambridge, MA: Harvard University Press.
Swedborg, R. 2005. *The Max Weber Dictionary: Key Words and Central Concepts*. Stanford, CA: Stanford University Press.
Tamir, Y. 1996. "The Quest for Identity." *Studies in Philosophy and Education* 15: 175–191.
Tarde, G. 1884. "Darwinisme naturel et darwinisme social." *Revue philosophique de la France et le l'étranger* 17: 607–637.
Tarde, G. [1890] 1921. *Les lois de l'imitation: Etude sociologique*. Paris: Alcan.
Taylor, C. 2003. "Sympathy." *A Philosophical Analysis*. London: Palgrave.
Thalos, M. 1999. "Degrees of Freedom: Towards a Systems Analysis of Decision." Journal of Political Philosophy 7: 453–77.
Theunissen, M. 1964. *Der Andere: Studien zur Sozialontologie der Gegenwart*. Berlin: de Gruyter.
Thomä, D. 1990. Die Zeit des Selbst und die Zeit danach: Zur Kritik der Textgeschichte Martin Heideggers 1910–1976. Frankfurt a.M.: Suhrkamp.
Thoreau, D. H. 1967. *Civil Disobedience*. New York: Twayne.
Tietz, U. 2002. *Die Grenzen des Wir: Eine Theorie der Gemeinschaft*. Frankfurt a. M.: Suhrkamp.
Tomasello, M. 1998. *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.
Tucker, A. W. [1950] 1980. "On Jargon: The Prisoner's Dilemma. A Two Person Dilemma." *UAMP* Journal 1: 101.

Tucker, A. 2001. "The Essence of Dissidence." *Graduate Faculty Philosophy Journal* 22/2: 59–78.
Tuomela, R. 1984. *A Theory of Social Action*. Dordrecht: D. Reidl.
Tuomela, R. 1991. "We Will Do It: An Analysis of Group-Intentions." *Philosophy and Phenomenological Research* 51: 249–277.
Tuomela, R. 1995. *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford, CA: Stanford University Press.
Tuomela, R. 2000. *Cooperation: A Philosophical Study*. Dordrecht: Kluwer.
Tuomela, R. 2002a. "Joint Intention and Commitment." In G. Meggle (ed.), *Social Facts and Collective Intentionality*. German Library of Sciences, Philosophical Research, Vol. 1. Frankfurt am Main: Hänsel-Hohenhausen, pp. 385–418.
Tuomela, R. 2002b. *The Philosophy of Social Practices. A Collective Acceptance View*. Cambridge: Cambridge University Press.
Tuomela, R. 2005. "We-Intentions Revisited." *Philosophical Studies* 125/3: 327–369.
Tuomela, R. 2007. *Philosophy of Sociality. The Shared Point of View*. New York: Oxford University Press.
Tuomela, R. and Tuomela, M. 2003a. "Acting as a Group Member and Collective Commitment." *Protosociology* 18: 7–65.
Tuomela, R. and Tuomela, M. 2003b. "Causal and Normative Group Responsibility." Unpublished manuscript.
Tuomela, R. and Miller, K. 1988. "We-Intentions." *Philosophical Studies* 53: 367–389.
Turner, S. P. 1999. "Searle's Social Reality." *History and Theory* 38: 216–228.
Turner, S. P. 2000. "Imitation or the Internalisation of Norms: Is Twentieth-Century Social Theory Based on the Wrong Choice?" In H. H. Kögeler and K. R. Stueber (eds.), *Empathy and Agency: The Problem of Understanding in the Social Sciences*. Boulder, CO: Westview, pp. 103–18.
Turner, S. P. 2004. "The New Collectivism." *History and Theory* 43/3: 386–399.
Tversky, A. and Kahnemann, D. 1986. "Rational Choice and the Framing of Decisions." *Journal of Business* 59: 251–278.
Udehn, L. 2001. Methodological Individualism: Background, History and Meaning. London: Routledge.
Velleman, J. D. 1989. *Practical Reflection*. Princeton, NJ: Princeton University Press.
Velleman, J. D. 1997. "How to Share an Intention." *Philosophy and Phenomenological Research* 57: 29–51.
Verbeek, B. 2002. *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*. Dordrecht: Kluwer.
Verbeek, B. 2007. "*Rational Self-Commitment*." In F. Peter and H. B. Schmid (eds.), Rationality and Commitment, Oxford: Oxford University Press, pp. 150–174.
Vermazen, B. 1993. Objects of Intention. *Philosophical Studies* 71: 223–265.
Volkelt, J. 1920. *Das ästhetische Bewußtsein*. Munich: C. H. Beck.
Waldenfels, B. 1996. "Sozialontologie auf sozialbiologischer Basis." *Philosophische Rundschau* 45: 97–112.
Walther, G. 1923. "Zur Ontologie der sozialen Gemeinschaften." In E. Husserl (ed.), *Jahrbuch für Philosophie und phänomenologische Forschung* VI. Halle: Niemeyer, pp. 1–158.
Weber, M. (ed.) [1919] 1921. "Politik als Beruf." In *Gesammelte Politische Schriften*. Marianne Weber (ed.). München: Drei Masken, pp. 396–450.
Weber, M. [1921] 1980. *Wirtschaft und Gesellschaft: Grundriß der verstehenden Soziologie*. Tübingen: J.C.B. Mohr.
Weber, M. 1922. *Gesammelte Aufsätze zur Wissenschaftslehre*. Marianne Weber (ed.). Tübingen: J.C.B. Mohr.
Weber, M. 1981. "Some Categories of Interpretive Sociology." Translated by E. Graber. The Sociological *Quarterly* 22/2: 151–180.
Weirich, P. 2004. "Economic Rationality." In A. R. Mele and P. Rawling (eds.), *The Oxford Handbook of Rationality*. Oxford: Oxford University Press, pp. 380–98.
Wilkins, B. 2002. "Joint Commitments." *The Journal of Ethics* 6: 145–155.

Williams, B. 1979. “Internal and External Reasons.” In R. Harrison (ed.), *Rational Action: Studies in Philosophy and Social Science*. Cambridge: Cambridge University Press, pp. 17–28.
Wojtyla, K. 1979. T*he Acting Person*. Analecta Husserliana, Vol. X. Dordrecht: D. Reidel.
Wrathall, M. and Malpas, J. (eds.) 2000. *Heidegger, Authenticity, and Modernity: Essays in Honor of Hubert L. Dreyfus, Vol. 1*. Cambridge, MA: Harvard University Press.
Wrong, D. H. 1992. “The Oversocialized Conception of Man in Modern Sociology.” In P. Hamilton (ed.), *Talcott Parsons: Critical Assessments, Vol. 2*. London: Routledge, pp. 211–224.
Wundt, W. 1900. *Völkerpsychologie: Eine Untersuchung der Entwicklungsgesetze von Sprache, Mythos und Sitte. Erster Band: Die Sprache. Erster Theil*. Leipzig: A. Kröner.

Index