Tzu-Wei Hung *Editor*

Communicative Action

Selected Papers of the 2013 IEAS Conference on Language and Action

Springer

Communicative Action

Tzu-Wei Hung
Editor

Communicative Action

Selected Papers of the 2013 IEAS Conference on Language and Action

Springer

Editor
Tzu-Wei Hung
Institute of European and American Studies
Academia Sinica
Taipei
Taiwan

ISBN 978-981-4585-83-5 ISBN 978-981-4585-84-2 (eBook)
DOI 10.1007/978-981-4585-84-2
Springer Singapore Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014936113

Printed on acid-free paper

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

Preface

This book focuses on the relationship between action and language. Despite intensive debates over action and language, few studies have examined how they are related and their shared underlying mechanisms. Some researchers claim that language is a special and highly structural case of action; that sensorimotor circuits form a cortical basis for language, and that language processing can be accounted for by sensorimotor interactions. Hence, the extent to which a mechanism for processing actions also facilitates processing language is an interesting question.

This book aims to foster a conversation among interdisciplinary scholars interested in unpicking the relationship between these two significant human capacities. This book is written for readers from different academic backgrounds—from graduate students to established academics, and readers will benefit from the diverse perspectives and extensive discussions of relevant issues.

Earlier versions of the essays in this book were presented at the 2013 IEAS Conference on Language and Action, held from September 17 to September 18, in Taipei, Taiwan. For financial and administrative support, I thank the Institute of European and American Studies, Academia Sinica (Chyong-Fang Ko, Director) and its Philosophy Group (especially Jih-Ching Ho, Wan-Chuan Fang, Norman Y. Teng, Timothy Joseph Lane). Following peer review, only some of the conference papers presented were selected for inclusion here, and I thank the referees for their work. I regret that I had to turn down several quality papers due to space limitations. Finally, I thank the editorial assistants, Yu-Tin Lin, Chih-Wei Wu, Kuei-Feng Hu, and Yi-Hsin Lai, for making this book possible.

Institute of European and American Studies
Academia Sinica
Taipei, Taiwan

Tzu-Wei Hung

Contents

Part I Language in Communication

Part II Action and Bodily Awareness

Part III Sensorimotor Interaction and Language Acquisition

Contributors

Ray Buchanan Department of Philosophy, University of Texas at Austin, Austin, TX, USA

Hsiang-Yun Chen Department of Philosophy, Centenary College of Louisiana, Shreveport, LA, USA

Tzu-Wei Hung Institute of European and American Studies, Academia Sinica, Taipei, Taiwan

Alistair Knott Department of Computer Science, University of Otago, Dunedin, New Zealand

Timothy Lane Taipei Medical University, Institute of Humanities in Medicine and Shuang Ho Hospital, Brain and Consciousness Research Center

Academia Sinica, Institute of European and American Studies

National Chengchi University, Research Center for Mind, Brain, and Learning

Rory Madden Department of Philosophy, University College London, London, UK

Hong Yu Wong Werner Reichardt Centre for Integrative Neuroscience, University of Tübingen, Tübingen, Germany

Syraya Chin-mu Yang Department of Philosophy, National Taiwan University, Taipei, Taiwan

Part I
Language in Communication

Names, Descriptions, and Assertion

Ray Buchanan

Abstract According to Millian Descriptivism, while the semantic content of a linguistically simple proper name is just its referent, we often use sentences containing such expressions "to make assertions…that are, in part, descriptive" (Soames 2008). Against this view, I show, following Ted Sider and David Braun, that simple sentences containing names are never used to assert descriptively enriched propositions. In addition, I offer a diagnosis as to where the argument for Millian Descriptivism goes wrong. Once we appreciate the distinctive way in which this account fails, we can better appreciate the very modest role that associated descriptive information plays in the pragmatics of proper names.

According to the traditional descriptivist theory, the semantic content of a proper name is given by a definite description (or cluster of descriptive information) that speakers associate with it; the name referring to whoever, or whatever, uniquely satisfies that descriptive information. As against this view, Kripke famously argued that, (a) speakers do not typically, and need not ever, associate uniquely identifying descriptive information with the names with which they are competent and (b) even in that rare case in which a speaker does have uniquely identifying descriptive information in her possession, it still does not follow that her use of the name refers to the unique entity that satisfies that information. For these reasons, as well as equally familiar Kripkean considerations concerning the rigidity of names, few theorists these days are sympathetic to the traditional descriptivist account.

Kripke's arguments gave rise to a widespread endorsement of Millianism—the view that the semantic contribution of a name is exhausted by its referent. But even if we agree with the Millian that the descriptive information associated with a name does not enter into the *semantic content* of an utterance containing it, this information might nevertheless play an essential role in the *pragmatics* of names. Indeed, in recent years, a number of theorists have argued in favor of a view we might call *Millian Descriptivism*—a view according to which proper names have a "Millian semantics," but "a partially descriptive pragmatics of assertion" (Soames 2008, p 283). Moreover, these theorists have argued that their favored pragmatic theory of names helps to explain some of the most well-known problems with Millian accounts of proper names.

R. Buchanan (✉)
Department of Philosophy, University of Texas at Austin, Austin, TX, USA
e-mail: raybuchanan@mail.utexas.edu

Tzu-Wei Hung (ed.), *Communicative Action,* DOI 10.1007/978-981-4585-84-2_1,

In what follows, I argue that Millian Descriptivism should be rejected. More specifically, I argue that the descriptive information we associate with a proper name no more enters into what we assert by our utterances involving it, than it does the literal, compositionally determined, semantic content thereof. As we will see, once we appreciate the distinctive way in which the Millian Descriptivist account fails, we can better appreciate the very modest role that associated descriptive information plays in the pragmatics of proper names.

1 Introducing Millian Descriptivism

In *Beyond Rigidity,* and a series of important subsequent essays, Scott Soames has argued that linguistically simple names have a "Millian semantics," but "a partially descriptive pragmatics of assertion" (Soames 2008, p. 283).[1] According to this *Millian Descriptivism*, while Millians are correct in holding the position that the semantic content of a simple name is just its referent, descriptivists are right in holding that we regularly use sentences containing such names "to make assertions, and express beliefs, that are, in part, descriptive" (Soames 2008, p. 283). Consider, for example (1)

(1) Bob Dylan is famous.

Qua *Millian* Descriptivists, these theorists hold that the sentence-type displayed in (1) semantically expresses the singular proposition (2):

(2) <Dylan, the property of being famous>

Qua Millian *Descriptivists*, however, these theorists emphasize that the semantic content of a sentence-type such as (1) will constrain, but not fully determine, what a sincere, competent speaker might *assert* by a literal utterance thereof. Rather, a speaker literally uttering (1) will oftentimes assert, and be understood as asserting, various *descriptively enriched propositions* (hereafter, "d-propositions"). More specifically, a speaker might assert—and, in so doing, intend to convey and undertake a commitment to the truth of—various d-propositions of the form displayed in (3) by uttering (1)

$$(3)\ \left[\text{The } x: \text{ F}x \text{ and } x = \text{Dylan}\right]\left[\text{Famous }(x)\right],$$

[1] A "linguistically simple name" is one for which "there is little… descriptive information that a speaker must associate with the name (qua expression-type) to be a competent user of it" (Soames 2002, p. 53). Such names contrast with "partially descriptive," complex names, like "Chief Justice Roberts," or "Rahway, New Jersey," which are associated with "substantial descriptive information that must be grasped by any competent speaker who understands and is able to use them correctly" (Soames 2002, p. 53). In what follows, I will only be concerned with "simple" cases. See, for example, Soames (2002, pp. 86–89) for an interesting discussion of the semantic contents of partially descriptive names. See Soames (2005) for some significant, and plausible, revisions to the account of semantic content offered in *Beyond Rigidity*.

where “F” is some, or other, contextually relevant property such as *being the guy who wrote* “*Blowin’ in the Wind,*” etc. The semantic content of the sentence-type (1) is nevertheless exhausted by (2), as it is the information that is *invariantly* contributed by that sentence to what one asserts in any normal, literal utterance thereof.

One of the principal selling points for Millian Descriptivism is that it seems to offer those theorists sympathetic with traditional Millian accounts of content a plausible means for responding to some of the familiar worries for their view. As the traditional Millian would be the first to emphasize, many speakers might initially be “resistant” to think that, for example, (4) and (5) have precisely the same semantic content.

(4) Bob Dylan is Bob Dylan.
(5) Bob Dylan is Robert Zimmerman.

The Millian Descriptivist can plausibly claim that when we ask ordinary speakers if two sentences of their language mean the same (i.e., have the same semantic content), they typically do *not*, as Soames puts it:

> … focus on the question of whether *what is common* to that which is asserted and conveyed in all contexts involving competent speakers by utterances of the one sentence is the same as *what is common* to that which is asserted by utterances of the other sentence. Instead they focus on what *they* typically would use the sentences to assert and to convey in various contexts, or what information *they* typically would gather from assertive utterances of them (Soames 2008, p. 283).

If Soames is correct, many “anti-Millian” intuitions are not so much evidence against the traditional Millian’s claims concerning semantic content as they are actually arguments *in favor* of the Millian Descriptivist’s account of what we assert by sentences with those semantic contents, since there always remains this common core. Indeed, the Soames-inspired Millian might point to the distinction between semantic content and assertoric content in virtually *any* case in which there seems to be a felt “mismatch” between her theory’s predictions concerning the former, and ordinary, competent speakers’ judgments concerning the truth conditions of utterances with that semantic content.[2]

2 Assertion, Expression, and Descriptive Enrichments

When assessing Millian Descriptivism, it is important to appreciate that the proponent of this view is not merely claiming that a speaker who uses a proper name is providing evidence that she has various beliefs with d-propositions as their contents. That much should, I think, be uncontroversial. Note that in a typical communicative exchange between a speaker, S, and her audience, involving a proper name, *n*, there

[2] Like Jeff Speaks, I suspect that Millian Descriptivism is currently “the most popular Millian reply to Frege’s Puzzle” (2010, p. 202).

will be a numerous descriptive conditions $D_1, \ldots, D_n$, such that it is *common ground* between them, that S associates these conditions with *n*.[3] In such a case, when S utters a sentence of the form "*n* is G" she will typically be providing her audience excellent evidence that she believes numerous d-propositions[4]:

$$\left[\text{The } x:\ \text{D}_1 x \text{ and } x = n\right]\ \left[\text{G}(x)\right]$$
$$\left[\text{The } x:\ \text{D}_2 x \text{ and } x = n\right]\ \left[\text{G}(x)\right].$$

Suppose, for example, I utter (6) in a context in which it is common ground between us that we both associate the descriptive conditions of *being the chair of the Horticulture department, being an enthusiast of home-brewed beer, being the person who ruined our couch*, and so on, with the name "Bobby Chantrelle"

(6) Bobby Chantrelle is coming to dinner.

By uttering (6), I will be giving you evidence that, among other things, I believe *that the chair of the Horticulture department, B.C., is coming to dinner*, etc. In this sense, I will be providing evidence of various beliefs of mine with d-propositions as their contents, and "weakly expressing" those d-propositions to you, even in those cases in which I do not intend to convey—much less *assert*—those propositions.

When a speaker genuinely intends to convey, and assert, some particular proposition by her utterance, her hearer must entertain that proposition if she is to successfully understand that utterance; not so for propositions that are merely weakly expressed. Unless I specifically intend to convey, say, (7) by uttering (6), you do not need to entertain that proposition in order to understand my utterance:

$$(7)\ \left[\text{The } x:\ \text{Ruined} - \text{our} - \text{couch}(x) \text{ and } x = \text{B.C.}\right]\ \left[\text{Coming} - \text{over} - \text{tonight}(x)\right].$$

Though you might reasonably infer that I believe (7) on the basis of my utterance,other facts you know of me, and the common ground, you will not have misunderstood my utterance should you fail to make that inference. Unless you take me to have intended to convey (7), your beliefs regarding the truth or falsity of (7) will be all but irrelevant to your beliefs regarding the truth conditions (and truth value) of my utterance of (6). A d-proposition that a speaker weakly expresses but does not actually intend to convey is no part of what she asserts, or what her audience who understands will take her to have asserted; it is, at best, a communicative by-product.[5]

[3] Let us say that S *associates* a descriptive condition D with *n* just in case S would, on competent, sincere, reflection, assent to "D(*n*)."

[4] Here, and in what follows, we can treat those cases in which it is common ground that *n*'s being G entails (or, makes highly probable) that *n* does not have the relevant D-property as "atypical."

[5] Once we appreciate that many of the effects of a speaker's assertion on the common ground are communicative by-product in the foregoing sense, we should resist any view on which the content of an assertion is simply read off from its "update" effects on the context of utterance. For example, any view that identifies the content of an assertion with the set of worlds compatible with the semantic content of the speaker's utterance and the common ground between her and her audience will fail to distinguish what the speaker actually asserted, and a by-product thereof.

There are other ways of reinforcing the point that unintended d-propositions are no part of what is asserted. If, for example, I literally utter "Bob Dylan played at Woodstock," I will be giving you good evidence that I believe that the famous rock star, Dylan, played at Woodstock; that the songwriter, Dylan, played at Woodstock, and numerous other d-propositions. But note that unless it is obvious that I have intended to convey such a d-proposition, you cannot target it for direct denial, or affirmation. For example, even if you disagree with me regarding whether Dylan is famous, wrote his own music, etc., you cannot felicitously deny or affirm the corresponding d-propositions by simply saying "No: that is wrong/false/incorrect. Dylan isn't famous/a rock singer/etc." Unintended, d-propositions are not typically *at-issue*, or easily available for agreement or disagreement, in the way we expect asserted content to be.

At any rate, while we can agree that speakers regularly weakly express d-propositions, admitting this much falls far short of vindicating the Millian Descriptivist claim that speakers regularly *assert* d-propositions, where this (minimally) requires both (a) *intending to convey* and (b) *intending to commit to the truth* of those propositions. In fact, Soames would agree as well, since he thinks that "for *p* to be asserted by an utterance of a sentence, it is not enough that conversational participants be in possession of information which, together with the speaker's utterance, might, after long or careful consideration, support an inference to *p*" (2002, p. 79). Rather, he thinks, we should require that "the speaker must know and intend that his hearers will take him to be committed to *p* on the basis of his assertive utterance, and the speaker must know and intend that the hearers are in a position to recognize this intention of the speaker" (Soames 2002, p. 80). Hence, in order to assess Millian Descriptivism, we should focus on the account's predictions in those cases in which it *is* obvious that conditions (a) and (b) are met. More specifically, we should ask whether in those cases in which a speaker utters a sentence containing a name intending to convey, and to commit to, a d-proposition *p*, it is plausible to claim that the speaker *asserted p*?

3 A Test Case

Let us turn then to a case in which the Millian Descriptivist will claim that a speaker asserts a particular d-proposition; a case in which the speaker manifestly intends to convey that proposition, and is recognized as having so intended. Suppose that Glenn believes that our neighbor, Freddy Morrell, is the former keyboard player of a famous 1980s band called "The Shrooms." Further, suppose that Glenn has, on numerous occasions, told us of this, believes that we believe this too, etc. Glenn associates the descriptive condition of being the former keyboardist of The Shrooms with the name "Freddy Morrell" and thinks that I do the same. As against this conversational background, suppose that while discussing an upcoming 1980s-themed fundraiser I am hosting for a local charity, I say to Glenn "I just don't know of any celebrities to invite that might impress the donors." He responds by uttering (8),

(8) Freddy Morrell is in town.

In this case, Glenn clearly intends to communicate both the d-proposition in (9), and that I can, and should, invite Freddy:

(9) $$\left[\text{The } x:\ \text{Former}-\text{Keyboardist}-\text{of}-\text{The}-\text{Shrooms}(x) \text{ and } x = \text{F.M.}\right]$$
$$\left[\text{Is}-\text{in}-\text{town }(x)\right].$$

Suppose I recognize that Glenn intends to convey both of these propositions as well as a true singular, unenriched proposition that the Millian Descriptivist would take to be the semantic content of (8). I submit that if there were ever a case in which a d-proposition is asserted by a simple sentence involving a proper name, this would be it.

As evidence in favor of the claim that Glenn asserted (9), by uttering (8), the Millian Descriptivist will point out how extremely natural and appropriate it would be for me to report what Glenn asserted or said by uttering (8), as in (10) (suppose you have just asked me if there are any local rock stars to invite to our party):

(10) Glenn asserted that the ex-keyboardist of the Shrooms, Morrell, is in town.

If we follow Soames in (a) thinking of assertion as "the most general and inclusive speech act of a set of closely related speech acts," including *saying that p, stating that p, claiming p*, and *telling H that p*, and (b) we assume that (10) is true just in case the relevant that-clause specifies something Glenn asserted it follows from (10) that Glenn indeed asserted a d-proposition [i.e. (9)]. As the Millian Descriptivist will emphasize, a report such as (10) sounds perfectly natural and appropriate in the described scenario, and the simplest explanation of why this is so is that (10) indeed truly characterizes what Glenn asserted (Soames 2002, pp. 73–77). Hence, absent any independent, compelling motivation for thinking that the report in (10) is not true, it might seem that Glenn indeed asserted the relevant d-proposition. (More on such reports anon.)

The Millian Descriptivist will argue that cases such as the foregoing are far from anomalous—speakers *regularly* assert, and are understood as having asserted, enriched by uttering simple sentences involving names.[6] Indeed, by her lights, it should be no more surprising that Glenn might assert (9) by uttering (8), than it is that I might assert that Mary got drunk and *then* drove home by uttering (11), or that Bill and Tom are married *to each other*, by uttering (12):

(11) Mary got drunk and drove home.
(12) Bill and Tom are married.

Once we appreciate that—quite generally—the semantic content of a speaker's utterance need not exhaust what she asserted by making that utterance, the Millian Descriptivist's claim that we sometimes assert d-propositions by our utterances involving names should seem considerably less surprising.

[6] Gleakos (2011) goes so far as to say that speakers *always* assert d-propositions by literal utterances containing proper names. According to her, the ubiquity of asserting d-propositions by our utterances involving names suggests that, contra Soames (2002), such propositions are also the semantic contents of these utterances.

4 The Case Against Millian Descriptivism

The foregoing case involving Glenn's utterance of (8) illustrates that, at least sometimes, speakers indeed intend to convey d-propositions by their utterances involving names. But even in those cases in which a speaker utters a sentence containing a name intending to convey, and to commit to, a d-proposition *p*, is it plausible to claim that the speaker *asserted p*?

In order to see how we might answer this question, first consider a principle I call (Assertion):

(Assertion) If S asserts that *p* by uttering *u*, and H understands *u*, but believes *not-p* at the time of utterance, then H will either (a) judge S to have asserted something false by uttering *u*, or else (b) change her mind on whether *p*, provided that she recognizes the inconsistency of *p* and *not-p*.

Though there is considerable disagreement among theorists on how, exactly, we should best understand the speech act of assertion, the foregoing principle will be (or at least should be) accepted by all parties. Though (Assertion) might need refinements to deal with tricky cases, it should (I hope) strike you as platitudinous. Three comments concerning this principle are in order: first, if we follow Soames in thinking of assertion as "the most general and inclusive speech act of a set of closely related speech acts" including *saying that p, stating that p, claiming p*, and *telling H that p*, we should, correspondingly allow that H's judgment in (a) might take the form of "S said/stated/claimed/told me something false by uttering *u*" (Soames 2002, p. 57). Second, for present purposes, we can leave condition (b) vague so as to allow that H might "change her mind" by coming to withhold belief in *p* (rather than come to believe *p*) as a result of S's assertion. Third, notice that (Assertion) should be attractive to the Millian Descriptivist; should the very minimal conditions in this principle fail, it is (at best) unclear how they could plausibly link facts about what a speaker asserts to our judgments regarding the truth conditions of utterances involving simple names in the way needed to "explain away" anti-Millian intuitions regarding semantic content.

Now with (Assertion) in mind, let us reconsider our example of Glenn's utterance of (8). Suppose the facts of the case are exactly as before except that we (Glenn's audience) both know that Freddy is *not* the ex-keyboardist of the Shrooms, and that, consequently, Glenn's belief to the contrary is actually *false*. (Imagine that though he has talked to us about this issue on many occasions, neither of us have ever had the heart to tell him that he was wrong on this score). But, as before, let us suppose that we recognize that Glenn nevertheless manifestly intends to convey (9) by uttering (8),

(8) Freddy Morrell is in town.

(9) $\left[\text{The } x\text{: Former} - \text{Keyboardist} - \text{of} - \text{The} - \text{Shrooms}(x) \text{ and } x = \text{F.M.}\right]$
$\left[\text{Is} - \text{in} - \text{town } (x)\right].$

Now, in this scenario, did Glenn assert (9) by uttering (8)? If (Assertion) is correct, no. While I fully recognize that Glenn intends to convey (9) by his utterance and actively believe the negation of that proposition, I have no inclination whatsoever to judge that Glenn asserted something false by uttering (8), nor do I change my mind on whether Freddy once played keyboards in the band, The Shrooms. Moreover, I strongly suspect that you would agree. Hence, from (Assertion) we should conclude that Glenn did not assert (9).

The foregoing case illustrates a point originally due to Braun and Sider (2006)—namely, that Kripke's famous "semantic argument" against classical descriptivist account of proper names seems to arise for the Millian Descriptivist, as well (Kripke 1980, pp. 83–87). Even in those cases in which it is obvious that a speaker is intending to convey a false description-theoretic proposition *p* by an utterance containing a proper name, we do not count the falsity of *p* as relevant to the truth conditions of her utterance.

This fact is as much a problem for the Millian Descriptivist account of asserted content, as it is for the classical descriptivist's account of the semantic content of such utterances.

The foregoing point is not limited to simple sentences involving names; it seems to be equally problematic for propositional attitude reports. Consider an example from Braun and Sider (2006, p. 672). Suppose that the host of a mathematics conference introduces Kurt Gödel to the audience as follows: "We are pleased to have the person who proved the incompleteness of arithmetic with us today. Prof. Gödel will speak on logic." Further suppose that (a) Smith and Jones are late for the lecture and *only* hear the host say "Prof. Gödel will speak on logic," and that (b) both believe—and wrongly take everyone else to believe—that Gödel is, in fact, an imposter who stole the incompleteness proof. Smith looks to Jones and whispers:

> Gödel stole the incompleteness proof from Schmidt! I really doubt he'll have the nerve to give a talk on logic. Surely he'll talk about something else. Still, the host believes that Professor Gödel will speak on logic. (Braun and Sider 2006, p. 672)

Now consider Smith's utterance of (13) in the foregoing dialogue:

(13) The host believes that Prof. Gödel will speak on logic.

As Braun and Sider point out, presumably the Millian Descriptivist is committed to claiming that in by uttering (13); in this scenario, Smith asserted that *the host believes that Prof. Gödel, who stole the incompleteness proof from Schmidt, will speak on logic*.[7] But note that the falsity of this proposition not withstanding, it is implausible that Smith asserted anything false by uttering (13). There is, as Braun and Sider put it, simply "no whiff of doubt" regarding the truth of the belief-report.

[7] Both Braun and Sider, and Soames, only discuss cases in which the appositive clause in the relevant belief report seems to reflect an aspect of how the agent—here, the host—is thinking of the object his belief concerns. In some cases, however, the sole function of such an embedded appositive clause is to help the belief ascribers's audience identify the object(s), or properties, that the agent's belief concerns (as in, for example, "Billy thinks that Glenn, who I told you about two days ago, likes chocolate"). In what follows, we will only deal with cases where it is plausible that the function of the appositive is to reflect an aspect of the content of the attitude the speaker is ascribing to the subject of the report.

In his response to Braun and Sider's discussion, Soames (2006) agrees that in the case just described, we do indeed have a "strong intuition" that Smith's belief report is true. He is skeptical, however, that the case is a genuine counterexample to the Millian Descriptivist position. According to Soames, speakers who use proper names in belief-reports, or elsewhere, typically assert numerous descriptively enriched propositions. In the case of (13), Soames claims that it is plausible that Smith asserted each of the following propositions, as well as "a number of related propositions" (Soames 2006, pp. 719–720),

The host believed that the day's guest, Gödel, would speak on logic.
The host believed that the person, Gödel, he was introducing would speak on logic.
The host believed that the man, Gödel, standing with him on stage would speak on logic.
The host believed that the logician, Gödel, would speak on logic.
The host believed that the well-known, Professor Gödel, would speak on logic.

In the case of (13), Soames claims that this "wealth of obvious enrichments produces an avalanche of truths" (Soames 2006, pp. 719–720). On the basis of this observation, Soames offers the following response to Braun and Sider:

> First, in considering what someone said we often focus on subpart of the whole of what was asserted: hence it is possible that this avalanche [of true d-propositions in the case of (13)] might mask the assertion of something false. Second, and I believe more significant for this example, our decisions about what descriptive enrichments should be credited to a speaker in determining his assertions may be guided, in part, by considerations of charity… when there are several obvious, relatively simple and straightforward truths the assertion of which may be credited to a speaker's remark, we may resist adding what we know to be a clear falsehood to the list, unless something about the discourse, or broader context of utterance, makes the addition unavoidable. (2006, p. 720; bracketed material mine)

Unfortunately, I do not think either of these considerations ultimately helps the Millian Descriptivist evade Braun and Sider's worry.

First, I doubt that our anti-Millian Descriptivist intuitions can always be explained by citing an "avalanche" of true asserted enrichments masking the presence of the problematic false enrichment. Suppose that you are in an unfamiliar neighborhood, and you approach a stranger, Tim, and ask whether there are any mechanics nearby. Tim responds by uttering (14),

(14) Tug McGee lives on Elmwood Drive.

Now, in this situation, we can stipulate that Tim only intends to convey a singular proposition concerning Tug, to the effect that *he* lives on Elmwood Drive, and the corresponding d-proposition containing an enrichment of that singular proposition with the property of being a mechanic. In this case, there is no avalanche of true d-propositions to appeal to in this case. Nevertheless, I submit that here, as before, we would judge that even if Tug is not a car mechanic, Tim did not assert or claim or tell you something false by uttering (14). No doubt, if Tim knows that Tug is not a mechanic he will be blameworthy for having intentionally misled you, but this need not be because he *asserted* the false enriched proposition—even a speaker who knowingly conversationally implicates a false proposition is guilty for having misled (more on this in a bit). Indeed, in some cases, a speaker can even be held responsible for weakly expressing a proposition he believes to be false.

Second, it is unclear (to me, at any rate) what Soames could have in mind by "the discourse, or broader context of utterance" making an enrichment "unavoidable" such that it is of help with regard to either the case of Glenn's utterance of (8), or Smith's utterance of (13). After all, in each of these cases it is obvious to both the speaker, and her audience, that she intends to convey, and intends to commit to, the truth of the relevant (false) enriched proposition. In the case of (8), for example, we can simply *stipulate* that it is mutually obvious to both of us and Glenn and that his utterance of (8) is an informative, relevant, answer to my query concerning who we should invite to the fundraiser on the condition that he intended to convey (9) thereby. Even though the relevant, false d-proposition is seemingly "unavoidable" in such a case, the problem remains—we do not judge Glenn to have asserted *anything* false, even though the relevant d-proposition [(9)] is false.

Third, I am doubtful that considerations of charity are of much help here. For the sake of argument, let us assume that, other things being equal, we invariably seek to minimize the number of obvious falsehoods that we take speakers to have actually asserted. For that matter, suppose that we *never* take speakers as having asserted enriched propositions that are obviously false. Even if we are charitable in this very thoroughgoing way, we can simply restate Braun and Sider's worry counterfactually: pick any case you like in which you successfully recognize that a speaker utters a sentence of the form in (15) intending to convey both a proposition concerning the referent of "*n*" to the effect that *it* is G, as well as some, or other, *true* d-proposition [The x: Dx and $x = n$] [G(x)]:

(15) *n* is G.

Now ask yourself whether it is plausible that the speaker would have asserted or claimed, etc. something false by uttering (14) *were* it to turn out that though the referent of "*n*" is G, it is not *D*. I submit that it is not. But since this counterfactual judgment does not itself require our taking the speaker to have asserted anything obviously false by her utterance of (15), it is doubtful the problem for the Millian Descriptivist can be explained by appeal to charity.[8]

Pending some alternative response to Braun and Sider's observation, I submit that we should reject Millian Descriptivism.

5 Descriptive Enrichment and the Pragmatics of Proper Names

Let us take inventory. The Millian Descriptivist claims that "we often *use* sentences containing (linguistically simple names) to make assertions, and express beliefs, that are, in part, descriptive" (Soames 2008, p. 283). We can agree with the Millian Descriptivist that, in some sense, we regularly "*express beliefs* that are, in part

[8] Alternatively, we could just ask whether the speaker asserted anything that *entails* that the referent of "*n*" is D, as well, so as to completely avoid the issue of the truth or falsity of the relevant enrichment.

descriptive” (italics mine) by using simple names. In virtually any use of a name we “weakly express” numerous such beliefs; and, sometimes, we even intend to convey, and are recognized as having intended to convey, beliefs with d-contents. The Millian Descriptivist is, however, mistaken in claiming that by using names in this way, we make *assertions* that are, in part, descriptive. As we have seen, even in those cases most congenial to the Millian Descriptivist—cases in which the speaker clearly intends to convey a d-proposition—it is implausible to claim that a descriptively enriched proposition was asserted.

But if Millian Descriptivism is false, then (a) what *do* we assert by our utterances of simple sentences involving names and (b) what is the status of the d-propositions in cases such as Glenn’s utterance of (8); that is, cases in which the speaker manifestly intends to convey a d-proposition? Since I suspect that my favored answers to these questions will be somewhat unsurprising in light of the foregoing critical discussion of the Millian Descriptivist, I will be brief.

Setting aside the very special case of empty names, I am sympathetic to the view that when a speaker literally utters a simple sentence of the form “*n* is G” what she asserts—and all that she asserts—is a singular, Millian proposition concerning the referent of “*n*” to the effect that *it* is G.[9] For example, returning to our original example concerning (1), I hold that the content of the act of assertion is the singular proposition in (2),

(1) Bob Dylan is famous.
(2) <Dylan, the property of being famous>

Of course, a speaker who literally utters (1) in order to assert (2) might also *conversationally implicate* numerous other propositions thereby. Indeed, in each of the cases we have discussed in which the speaker clearly intends to convey a d-proposition ((8), (13), and (14)), the relevant enrichment is plausibly part of what the speaker implicates *by* asserting a singular proposition in the specific manner she did.[10] But while a descriptively enriched proposition might be among the things a speaker intends to communicate by her utterance involving a name, such a proposition will nevertheless be (at best) something she indirectly means, and suggests, by asserting what she did—a proposition that her audience will take her to have meant by her utterance if they are to preserve the presumption that she was being conversationally cooperative. For example, Glenn’s utterance of (8) constitutes an informative, cooperative response to the question under discussion in part, because he meant the d-proposition (9). One very pleasing consequence of this diagnosis is that we should expect that our evaluation of the truth, or falsity, of a speaker-meant d-proposition will be all but irrelevant to our evaluation of the truth conditions of

[9] I am sympathetic to the proposal of Braun (1993) on which both the semantic content of an utterance containing a nonreferring proper names is “gappy proposition.” See Buchanan (2010, 2013, for an attempt to make sense of gappy propositions as the contents of our assertions.

[10] I take each of these cases to crucially involve both the maxim of manner and the maxim of relevance.

what the speaker asserted by the relevant utterance. That is, we should expect the results from Section 4 that looked so problematic for the Millian Descriptivist.

In short, I am sympathetic to the traditional view according to which the things we assert by literal utterances of simple sentences involving proper names are singular, Millian Propositions, allowing that sometimes we might also conversationally implicate any number of other propositions—including, in some cases, descriptive enrichments—by such utterances.

In endorsing this "old school" variety of Millianism, I do not mean to suggest that the descriptive information we associate with names might not play some more limited, modest role in the pragmatics of names. Perhaps it does. Note that in virtually any case in which a speaker literally uses a proper name "*n,*" she will intend for her audience to infer who, or what, she is referring to (in part) by trying to find some object in the common ground that bears that name. But in those cases in which there is more than one object in the common ground that bears the relevant name, she might have to rely on her audience having *further* information—including shared descriptive information associated with the "*n*"—in their wherewithal which will help put them into a position to infer which so-called thing she is intending to refer to. This point should be intuitive—think of what justifies you in expecting that your audience will take you as referring to one bearer of, say, the name "Bob" rather than another, in a particular context of utterance. If I utter "Bob is playing a show tonight," you might come to recognize which bearer of that name I am referring to—say, Bob Dylan—at least in part as a result of your knowing that it is common ground between us that a certain bearer of the name "Bob" has *the property of being a famous singer or songwriter*. Here, as before, this descriptive information is *not* plausibly part of what I assert by uttering "Bob is playing a show tonight" (note, e.g., I will not have asserted something false should it turn out that Dylan did not write his own songs). Rather, it is merely information that I intend for my audience to rely on in coming to correctly identify what it is that I asserted.

Earlier we saw that the descriptive information we associate with a proper name (or take others to associate with it) might sometimes enter into the propositions that we "weakly express," or conversationally implicate, by utterances of simple sentences containing it. In light of the preceding paragraph, we can add that such associated descriptive information might also sometimes figure in the information we intend our audiences to use in inferring what we have asserted by our utterance involving the relevant name. Crucially, however, that information it is never part of what we *assert* by such utterances.

Acknowledgments I would like to thank the organizers and all of the participants of the 2013 Conference on Language and Action at the Institute of European and American Studies at Academia Sinica for extremely helpful feedback on a presentation of an earlier draft of this chapter. I would also like to thank Hsiang-Yun Chen, Josh Dever, Sinan Dogramaci, Matt Evans, Hans Kamp, Bryan Pickel, Gary Ostertag, and David Sosa for discussion of issues in the theory of reference relating directly to this material. I owe a very special thanks to Gary for comments on an earlier draft that significantly improved the final chapter. All of the normal qualifications and disclaimers are in order.

References

Braun, D. (1993). Empty names. *Noûs, 27,* 449–469.

Braun, D., & Sider, T. (2006). Kripke's revenge. *Philosophical Studies, 128,* 669–682.

Buchanan, R. (2010). A puzzle about meaning and commincation. *Noûs, 210,* 340–371.

Buchanan, R. (2013). Reference, understanding, and communication. *Australasian Journal of Philosophy*. doi:10.1080/00048402.2013.788526.

Gleakos, S. (2011). The propositions we assert. *Acta Analytica, 26,* 165–173.

Kripke, S. (1980). *Naming and necessity*. Cambridge: Harvard University Press.

Soames, S. (2002). *Beyond rigidity: The unfinished semantic agenda of naming and necessity*. Oxford: Oxford University Press.

Soames, S. (2005). Naming and asserting. In Z. Szabó (Ed.), *Semantics versus pragmatics* (pp. 356–382). New York: Oxford University Press.

Soames, S. (2006). Reply to critics. *Philosophical Studies, 128,* 711–738.

Soames, S. (2008). *The gap between meaning and assertion: Why what we literally say often differs from what our words literally mean. In Philosophical essays: Vol. 1. Natural language, what it means & how we use it* (pp. 278–297). Princeton: Princeton University Press.

Speaks, J. (2010). Millian descriptivism defended. *Philosophical Studies, 149,* 201–208.

Indefinites in Action

Hsiang-Yun Chen

Abstract Karen Lewis (Philos Stud, 158:313–342, 2012) argues that recognizing the importance of plans helps settle a debate regarding the semantics and pragmatics of indefinites. More specifically, Lewis argues against the dynamic approach (e.g., Kamp (In Groenendijk et al., Formal Methods in the Study of Language, pp. 277–322, Mathematics Center, Amsterdam, 1981), Heim (The semantics of definite and indefinite noun phrases, University of Massachusetts at Amherst, 1982), Groenendijk and Stokhof (Linguist Philos, 14:39–100, 1991), Kamp and Reyle (From Discourse to Logic, Kluwer, Dordrecht, 1993), and Asher and Lascarides (Logics of Conversation, Cambridge University Press, Cambridge, 2003)), according to which indefinite expressions are subject to a semantic Novelty condition. Drawing on data of the so-called summary uses, she claims that Novelty is best analyzed as a pragmatic, cancelable implicature. This chapter throws significant doubt on Lewis' analysis. Not only is her objection in large part a misreading of dynamic semantics, but the proposed pragmatic account offers no real explanation of even the alleged counterexamples. Once we consider a wider range of linguistic phenomena involving indefinites, the verdict is on the side of the dynamic approach.

1 Introduction

Here is a widely endorsed picture of the meaning of linguistic expressions. The semantic content of a sentence is its truth conditions, and the semantic content of subsentential expressions are their contributions to the truth conditions of the sentences in which they are embedded. Following Russell (1905), indefinite expressions, i.e., expressions of the form "a F," are semantically equivalent to existential quantification.

It is hard to ignore, however, that indefinites play a dual function in actual practice: They not only assert existence, but introduce an element that can figure in subsequent discourse.

H.-Y. Chen (✉)
Department of Philosophy, Centenary College of Louisiana, Shreveport, LA, USA
e-mail: hsiangyun@gmail.com

Tzu-Wei Hung (ed.), *Communicative Action,* DOI 10.1007/978-981-4585-84-2_2,

Consider:

(1) Mary completed a paper on radical skepticism. She has just submitted it to a journal for review.
(2) Every man who has a daughter adores her.

Intuitively, we have no problem understanding the pronouns as anaphoric, yet traditional, Russellian, static semantic theories have few resources to explain how indefinites can license anaphora beyond their syntactic binding scope.

In contrast, *dynamic* semantics takes the meaning of a sentence as its *context change potential* (CCP), and the meaning of subsentential expressions is their contribution to the CCP of the whole. This by no means suggests that truth conditions are not important, but that in order to fully capture what goes on in linguistic communication, one needs to embrace a broader notion of meaning and keep track of more fine-grained information. In particular, one needs to keep a record of "things being talked about" or the "objects under discussion" in a conversation. The discourse interlocutors' task in understanding what is being said in the course of a conversation thus consists in (a) cataloguing an inventory of *discourse referents* as well as (b) altering the information associated with them as the discourse unfolds, where strictly speaking only the latter is truth evaluable.

Dynamic theories make the aforementioned dual function of indefinites explicit by offering a nonquantificational analysis: the CCP of indefinites is the introduction of a new "discourse referent" in discourse representation structures (DRS, e.g., Kamp (1981) and Kamp and Ryle's (1993) discourse representation theory (DRT)), or the addition of a "file card" in file change semantics (FCS, e.g., Heim 1982, 1983). The existential quantification traditionally associated with indefinites is implicit: It is construed as part of the verification condition of a DRS or the satisfaction condition of updating a file with an utterance that contains an indefinite. By taking the Novelty Condition as the defining characteristic of indefinites, together with a formalism that allows for a wider binding scope, dynamic theories can successfully account for sentences like (1) and (2).

Recently, Lewis (2012) claims that the dynamic approach is mistaken, because Novelty cannot be semantic. She argues on the basis of what she calls the "summary uses" of indefinites that Novelty must be analyzed as a pragmatic, cancelable implicature. In addition, drawing heavily on the idea that planning and plan recognition are central to conversation and communication, she offers a neo-Gricean account that purportedly explains the uses of indefinites that conform to Novelty as well as those that do not.

This chapter aims to show that Lewis' analysis is misguided. The alleged counterexamples to the dynamic approach to indefinites, i.e., the so-called summary uses, are at best dubious and borderline illusionary. The rest of the chapter is structured as follows. Section 2 reviews Lewis' argument that Novelty cannot be semantic and sketches her pragmatic analysis. In Sect. 3, I show that Lewis' criticism of dynamic theories reviewed in Sect. 2 results from a confusion. Moreover, I demonstrate that Lewis' own pragmatic proposal not only rests on problematic bases, but falls short of explaining the diverse linguistic phenomena involving indefinites, including the very examples she puts forth. Section 4 discusses some general lessons from this dialectic.

2 A Pragmatic Account

The underlying thesis of Lewis' treatment of indefinites is that they are simply existential quantifier; indefinites have their traditional Russellian *semantic* content. She acknowledges that this simple static account does not suffice to explain the behavior of indefinites in discourse, but argues that what is needed is nothing more than a supplementary pragmatic story of the familiar Gricean kind. Specifically, a broadly Gricean story is required to spell out how indefinites are capable of (a) introducing a new "object under discussion" into the conversation and (b) licensing anaphora beyond their standard binding scope. Crucially, the claim is that not only can these two features, i.e., Novelty and Licensing, be accounted for in a pragmatic fashion, but that such a pragmatic picture is also empirically superior to the competing dynamic semantic theories.

The crucial evidence for the second point comes from examples such as the following:[1]

(3) a. A student walked into Sue's office and asked her about his exam.
b. Finally, a student needed her help!
(4) a. I went to see *Star Trek* on Sunday.
b. That's pretty much all I did all weekend: I saw a movie.
(5) a. We have this nail here.
b. Unfortunately, now we have a nail and no hammer.
(6) a. I went out to dinner with the woman from the bar last night.
b. Can you believe it—a woman went out to dinner with me!

Another example of this kind can be found in Gundel et al. (1993):

(7) a. Dr. Smith told me that exercise helps.
b. Since I heard it from a doctor, I'm inclined to believe it.[2]

In all these examples, the indefinite expressions in (b)—"a student," "a movie," "a nail," "a woman," and "a doctor"—do not pick out a new object in the discourse. Rather, their use is justified by an object previously mentioned in (a).[3] So, Lewis argues, (3) through (7) are not *introductory* uses; rather, they exemplify the *summary* uses of indefinites. More importantly, if novelty is a semantic feature of indefinites, it must be conventional, systematic, and cannot be overridden. That summary uses

[1] The following examples are from Lewis (2012) examples (6)–(9) on p. 318.

[2] Gundel et al. (1993, p. 296), example (49).

[3] The "antecedent" of the indefinites may be an indefinite (as in (3) and (7)), a proper name (as in (13)), a demonstrative (as in (5)), or a definite (as in (6)).

of indefinites are felicitous and robust and thus argues strongly against treating novelty as semantic.

Lewis then contends that a broadly Gricean pragmatic analysis of novelty is preferable, as it naturally explains the existence of both the introductory and summary uses. So long as novelty is treated as an implicature, there is no surprise that it is sometimes cancelable. Indeed, cancelability is often viewed as an indicator that the phenomenon in question is pragmatic rather than semantic. The real challenge, however, is the provision of a plausible and coherent pragmatic story. According to Lewis, recognizing planning as fundamental in conversation and communication is the key to such a pragmatic analysis.

Humans are essentially planning creatures. We are intelligent *actors* that inhabit complex, dynamic environments, which we manipulate in complex ways. One of the important ways that we connect and effect our environments, including other agents, is through language. From this perspective, "a well-run conversation is just like any other cooperative, rational activities" (Lewis 2012, p. 322). "A successful conversation [also] requires a coherent series of plans: not just what to talk about or how to answer a question under discussion, but also how an object under discussion relates to a question under discussion" (Lewis 2012, p. 323). In other words, interlocutors "do not make random, disconnected utterances."

Planning and plan recognition are clearly closely related to intending and intention recognition. Lewis maintains that thinking of plan recognition as central is compatible with and extends Gricean pragmatics, since it emphasizes, besides "what a speaker wants the interlocutors to believe (or understand, or presume)," "how the speaker wants to fit her contribution into the overall conversation." Moreover, the plan recognition framework provides a natural explanation of the fundamental interrelatedness between Grice's maxim of relation (i.e., be relevant) and the maxim of manner (i.e., be perspicuous). Acknowledging that a complete plan for a conversation is oftentimes not predetermined,[4] Lewis nevertheless argues that *local discourse plan* should be recognizable, as they are the driving forces of particular utterances. A local plan is recognizable partly because it connects to the overall discourse plan in a transparent way. Put differently, "recognizable, perspicuous plans go hand in hand with relevant utterances."

But how does this explain Novelty and the introductory uses of indefinites? Consider (8):

(8) a. A woman walked in.
    b. She looked gloomy.

Lewis' derivation goes like this. Semantically speaking, a sentence with an indefinite is simply a general, existential claim. By assumption, participants of a conversation are cooperative so that they only make relevant contributions to the conversation. Hence, the existential claim made in (8a) must be relevant to the conversational context and the overall discourse plan. Rational, cooperative conversation

[4] "[A] complete plan for a typical conversation is not decided upon beforehand, but the sort of plans we will be concerned with are speakers' short-term plans, which we can call local plans" (Lewis 2012, p. 323).

participants would ask themselves why the speaker makes this specific choice. More specifically, had the speaker wanted to talk about a woman already under discussion, she had less misleading ways to do so; a pronoun, definite description, or name would all be more appropriate. Therefore, using the indefinite "a woman" is indicative of a plan to convey information about a new woman under discussion. Furthermore, the use of an indefinite is frequently a marker of a plan to say something further about its referent, which accounts for the anaphoric pronoun in (8b).

In short, Lewis argues against the semantic analysis of Novelty and Licensing. She makes no objection against the file-card metaphor, however, so long as it is understood pragmatically. Tracking a conversation, or updating a conversational context, is a pragmatic process that involves plan recognition; the conversational context that interlocutors must keep track of is, at the very least, a stack of file cards, or a collection of the objects under discussion. From the speaker's point of view, the use of an indefinite is a perspicuous way to signal that a new object is being introduced into the conversation; the addressee grasps the speaker's communicative intention and understands the speaker's utterance as relevant to the overall discourse, resulting in the addition of a new card. Planning and plan recognition are not ad hoc; they are general reasoning mechanisms that are independently motivated. Taking them seriously as the underlying principles governing discourses makes the explicit coordination necessary for communication.

3 Weighing Between Semantics and Pragmatics

While I agree that planning and plan recognition are crucial in rational, cooperative activities, and linguistic communication should be no exception, I think Lewis' objection to the dynamic semantic theories and her own pragmatic treatment do not stand close examination. First, it strikes me that her criticism of the dynamic theories is largely a misreading. Once the thesis of the dynamic approach is properly understood, it is a plain illusion that the so-called summary uses of indefinites pass as counterexamples. Furthermore, Lewis' positive proposal lacks its claimed explanatory power. It hinges on dubious assumptions and does not adequately account for either the specific examples Lewis herself brings to spotlight or data involving indefinite expressions in general.

3.1 Is There Anything Wrong with the Semantic Approach?

To begin, as the dynamic theories conceive it, Novelty is not a matter of reference, or objects in the world (i.e., the model), but a constraint on the construction of the semantic representation of the utterance containing an indefinite. In Heim's FCS or Kamp's DRT, the CCP of an indefinite is the introduction of a new file card to the file or a new discourse referent to the DRS, where a file or a DRS is a theoretical, representational construct mediating between language and the world. Novelty simply leaves open whether distinct cards or discourse referents are mapped to the same or different objects in the model.

In dynamic theories, discourse reference and genuine reference are two distinct notions. Here are some quotes from Heim (1983):

> [D]iscourse referents behave in ways which it wouldn't make any sense to attribute to real referents: not only are there discourse referents for NPs that have no referents, but moreover, discourse referents may suddenly go out of existence, depending on certain properties of the utterance.
> [I]t is quite conceivable for there to be a file card that fails to describe a referent, or for two different file cards to happen to describe the same thing, or for file cards to be introduced into and be removed from the file, depending on what is getting uttered.[5]

In fact, with this distinction firmly in place, Heim discusses an example that bears much on our present discussion:

(9) John came, and so did Mary. *One of them* bought a cake.[6]

"One of them" is an indefinite noun phrase (NP), but clearly its referent, be it John or Mary, has been mentioned in the first part of (9). This, however, is not a violation of the Novelty condition. The prediction about "one of them" is simply that "its discourse referent must be new and must be distinct from the discourse referents of "John" and "Mary" in particular. There is no prediction about the reference of these three NPs, and we may consistently hold any assumption we please about those. In particular, we may assume that NPs with *discourse reference sometimes happen to coincide in reference* (my italics), and that [(9)], being a case of this kind, involves three discourse referents, but only two referents" (Heim 1983, p. 166).

As is clear from this example, a new discourse referent, or file card, does not entail a new individual. Judging from this light, examples of the summary uses are no challenges to the Novelty condition as dynamic theories depict it.

To be fair, Lewis is not completely unaware of this. She notes that "[i]t is important to note that novelty is not a matter of reference or denotation; no one claims that the object in the world that actually satisfies the indefinite description has to be new to the conversation. Novelty is the claim that, roughly, a speaker is talking about something that is novel for the purposes of the conversation." Also, "On these views [of dynamic semantics], the CCP of an indefinite description dictates that a novel representation of an object under discussion should be added to the context. This representation then provides a value for subsequent anaphoric pronouns. Objects under discussion are represented in the context by discourse-level entities, i.e., representations that are neither linguistic expressions nor objects in the world (or in a model)" (Lewis 2012, p. 316).

Discourse-level entities, whether they are called file cards or discourse referents, are merely representations of the objects, which are under discussion, in the world. However, when Lewis goes on to discuss the summary use and treats it as evidence against the novelty condition, she completely disregards the representational level.

[5] The quotes are from Heim (1983, pp. 166, 168), respectively.

[6] Heim (1983, p. 165).

She writes, "[…] the second sentence in each discourse contains an indefinite that intuitively continues talking about an object already under discussion." Also, "[…] [t]he most salient interpretation (if not the only interpretation) of [(3)] is that one and the same student walked into Sue's office and needed her help." (Emphasis mine)[7]

Ultimately, the problem here is that Lewis' challenge rests on a shifty notion of "object under discussion": At times she uses the term in its representational sense (as with respect to the introduction use), but at other times she uses it in the genuinely referential uses (as with respect to the summary use). So, the argument is fallacious because of the equivocation. When these two interpretations are carefully differentiated, as they should be, Lewis' argument is at best categorically mistaken.[8]

To be sure, what is interesting about (3) through (7) is that the file cards must have the same reference. Lewis briefly considers a potential response from the proponent of a dynamic semantic account that explores the *merging* of file cards: File cards may be merged when conversation participants realize that what were being treated as distinct objects under discussion are in fact satisfied by the same object in the world. She then criticizes that merging is ad hoc and unsatisfactory as it "saves a technical notion of novelty" by sacrificing the significance and explanatory power of the file-card metaphor. While I am not convinced that the merging process is ever needed, I am sympathetic to the concern of how contentful the Novelty constraint really is.

Still, treating the summary uses as a decisive reason against a semantic account strikes me as a hasty conclusion. I have two points to make on this score. First, note that in the examples of summary uses (i.e., (3)–(7)), expressions such as "finally," "can you believe it," and "since" play an important role. The minimal pair (10) and (11) provides a vivid illustration:

(10) a. A student walked into Sue's office and asked her about his exam.
b. A student needed her help!
(11) a. A student walked into Sue's office and asked her about his exam.
b. Finally, a student needed her help!

In the absence of "finally," the discourse in (10) allows for various interpretations; but in (11), there is no such flexibility. It is no longer ambiguous whether the two occurrences of "a student" pick out the same individual. The summary uses become natural only when there is a discourse particle that signals the discourse structure

[7] Lewis (2012, p. 316).

[8] Here is another way to block Lewis' argument. Consider a scenario where whenever a student needed Sue's help, he did not go to her but asked one of his classmates instead. In this case, the second occurrence of "a student" is still a summary use in the relevant sense, yet it no longer denotes "one and the same student" as the first occurrence of the indefinite noun phrase, i.e., the student that walked into Sue's office. Of course, such possibility is disastrous for Lewis' account, but provides further evidence that favors the dynamic analysis. I am grateful to Josh Dever for drawing my attention to this possibility.

by marking the rhetoric relation between sentences (a) and (b). This is in complete agreement with the predictions of the dynamic theories.[9]

Moreover, the effects these structural markers contribute to do not seem to be cancelable. Consider

(12) a. A student walked into Sue's office and asked her about his exam.
b. Finally, a student needed her help!
c. #But he is not the same student as the first one./#But they are not the same students.

What happens in (12) is that once the second occurrence of "a student" is interpreted as an instance of the summary use, that bit of information cannot be overridden no matter how the conversation further develops. If cancelability is a marker of pragmatics, then the difference in meaning that "finally" brings about to the overall discourse looks more like a semantic contribution.

The contrast between (10), (11), and (12) is evidence that indefinite expressions (e.g., "a student") and discourse particles (e.g., "finally") must interact in such a way that systematically constrains how the discourse can be interpreted. On the one hand, the two occurrences of "a student" need not pick out the same individual in the model in (10), though they must so in (11). On the other hand, the use of subsequent anaphora is highly regulated: While the speaker in (10) may carry on with the information that she is really talking about two distinct students, she cannot do so once the sentence that contains "finally" appears in the discourse, as (12) demonstrates. If the interplay between indefinites and markers of the rhetoric relations is limited at the pragmatic level only, however, it makes no sense why the summary uses cannot be retracted.

In short, Lewis' objection to the dynamic accounts is misleading. Once we recognize the status of file cards as theoretical, representational entities, as the dynamic theorists have it, examples involving the summary uses pose no real challenge. Even if these examples raise a question of the purpose of the Novelty constraint, they are no knock-down arguments against a semantic treatment. As a matter of fact, considerations of the interaction between indefinites and other parts of the discourse, particularly those signaling the conversational structure, favor such a treatment.

[9] Earlier dynamic theories, e.g., Kamp (1981); Heim (1982); and Kamp and Reyle (1993), only predict the ambiguity of (10) and strictly speaking do not fully explain (11). But more recent DRT-based theories, such as Asher and Lascarides' Segmented Discourse Representation Theory (SDRT), do account for the semantic contributions of discourse particles. Very roughly, discourse particles signal the rhetorical structures, e.g., elaboration, consequence, contrast, explanation, etc., so that the discourse relations place more constraints on the accessibility conditions in the DRT-style model theory. Since the primary aim of this chapter is to demonstrate that Lewis' objections to the dynamic approach are misguided, I will not discuss the full details of a complete explanation of (11) along the lines of SDRT.

3.2 What Is Not Right About the Pragmatic Approach?

Lewis' own analysis of the relevant phenomena is equally dissatisfying. Besides lacking crucial details regarding the nature of plans, her theory suffers from obvious counterexamples and does not even explain the data she herself raises to salience.

One fundamental difficulty with the kind of account Lewis proposes concerns the speaker's explicit denial of any discourse plan. Take

(13) a. I do not have any plan in telling you the following.
b. A student walked into Sue's office and asked her about his exam.
c. Finally, a student needed her help!

Despite the speaker's straightforward confession that she has no plan for the conversation, the addressee would engage in some plan recognition: The speaker's utterance of the indefinite "a student" in (13b) introduces into the conversational context a new file card no matter what. This raises the question of the nature of plans that the addressee is supposed to be able to recognize.

At one point, Lewis states that "speakers use and participants recognize maximally strategic plans." (2012, p 329) Taken as an unrestricted, empirical claim, this contention is plainly false as conversations are oftentimes random and extemporaneous. Her claim is more realistically viewed as an idealization or the goal of conversations. But what are maximally strategic plans and what makes them recognizable? One would expect an account that rests on the centrality of plans to address these fundamental questions. Yet Lewis says surprisingly little on either, and what she does say raises more worries.

In the artificial intelligence (AI) literature, planning is typically understood as "the process of formulating a program of action to achieve some specific goal" (Pollack 1992, p. 3). Given some initial conditions and the specification of a specific goal, the planning agent (or system) produces a series of actions whose execution will achieve that goal. I am not sure if this is the picture Lewis has in mind, for she wants to "remain neutral" on the nature of plan. She does, however, assert that her focus is on the speaker's short-term, or *local*, plans, which may be thought of as elements or subplans of an overall plan. Crucially, local plans should be recognizable, as they are the type of plans that "drive particular utterances."[10]

It strikes me that there is a puzzle regarding the connection between local plans and the overall discourse plan. On the one hand, Lewis admits that "a complete plan for a typical conversation is not decided upon beforehand." Yet according to her, a well-run conversation must be one where the local plans are maximally relevant and perspicuous with respect to the discourse plan. But if a complete plan is not established in the first place, it is unclear how local plans—the pragmatic import of subsentential, subdiscourse elements—can ever be judged as relevant and perspicuous. On the other hand, the problem that discourses such as (13) bring out is even more telling. In the sheer absence of an overall discourse plan, what maximally stra-

[10] The recognition of local plans allows "the participants to track the discourse, i.e., know what to expect will likely be a topic of conversation, an object under discussion, or a question being addressed" (Lewis 2012, p. 329).

tegic local plans can there be? To maintain the idea that conversation participants recognize maximally strategic local plans, Lewis would have to admit that these plans must, in general, be autonomous. But then it makes little sense to talk about local plans coming together and being relevant and perspicuous for the purpose of a conversation. Once again, the relation between local plans and the entire discourse becomes a mystery. Furthermore, as the denial of a complete discourse plan can be easily generalized, it is not helpful to counter the challenge by restricting the analysis to task-oriented dialogues. Doing so seriously reduces the significance of the theory and leaves the real problem unresolved.

Whether or not (13) is deviant, the general points it illustrates are clear enough. A theory of linguistic understanding and communication that builds upon planning and plan recognition is faced with two interconnected tasks of coordination. First, it must explain what makes the coordination between the speaker and the addressee possible. It must allow for the potential gap between (a) the speaker's possibly nonexistent plan, incomplete plan, or multiple plans and (b) whatever plan(s) the address is able to recognize. Second, it must explain what makes up a discourse plan. If its composition involves fine-grained levels of subplans, it must account for the contributions these subelements make to the global plan, and how this bears on the speaker's production and the addressee's understanding. There can be no dodging a precise explication of the nature of plans and what is it that makes plans recognizable, yet Lewis' proposal fails to adequately address these metaphysical and epistemological issues.

My second objection concerns the pragmatic analysis' inability to successfully account for the relevant linguistic phenomena, including both the summary uses and the introductory uses.

First, consider the following:

(14) a. A student walked into Sue's office and asked her about his exam.
b. Finally, a student needed her help!
(15) a. A student walked into Sue's office and asked her about his exam.
b. (?) Finally, some/at least one student needed her help!
(16) a. A student walked into Sue's office and asked her about his exam.
b. (?) Finally, he/the student/John needed her help!

Replacing the occurrence of "a student" in (14b) with other truth-conditionally equivalent phrases results in at least some difference in acceptability. We may consider two types of substitution: (a) substituting "a" with other indefinite expressions like "some" and "at least one"; (b) substituting "a student" with a *definite* expression—the pronoun "he," the definite description "the student," or a proper name. Even if type (a) substitution is marginally acceptable, type (b) substitution appears much worse.[11] However, it is not clear how the pragmatic analysis of indefinites can coherently explain these phenomena without being self-defeating. Here is a quote from Lewis:

[11] It seems to me that if (16) is to make sense at all, the discourse as a whole means something quite different from (14). The speaker must assume her addressee to have a much stronger degree of familiarity with the said individual.

> In the summary uses the existential, general meaning of the indefinite is emphasized. [(14b)] is appropriate to utter in a context in which Sue had been waiting and hoping for *some student or other* to need her help—she isn't happy or relieved because that particular student came to her office in need of help, but that some student at all needed her help. The speaker has a special reason to use an existential claim, since it *expresses something a definite expression cannot* (italics mine). If we replaced *a* with *the* in the summary uses, they would each convey something different, if they made sense at all. Since there is this special reason to use an indefinite and only an indefinite, we have reason to believe novelty won't be implicated. (2012, p. 332)

This remark is curious. Earlier on, Lewis argues that "[t]he most salient interpretation (if not the only interpretation) of [(14)] is that *on and the same* student walked into Sue's office and needed her help" (emphasis mine),[12] which strongly suggests that the summary use of indefinites is anaphoric. But here she explicitly states that what indefinites do in their summary uses is something definites cannot. She submits that while the introductory uses are meant to pick out an individual, the summary uses are *not* supposed to pick out any specific individual, though such uses are justified by one. In a nutshell, it is the purely existential, general meaning that underlies the summary use.

This strikes me as evidence that Lewis' analysis is inconsistent. On the one hand, she argues that summary uses undermine the Novelty condition because in such a use an indefinite is about an object already under discussion. Yet, if it is truly about *one and the same* individual, there is no explanation what distinguishes between definites and the summary uses of indefinites. On the other hand, the admission that the summary uses are not really about a specific individual vindicates the idea that they too introduce some new file cards into the conversational context, just like the introductory uses. However, the file cards triggered by the summary uses are special: They denote a concept or a category rather than individual instances thereof. This way, the summary uses of an indefinite are different from the introductory uses in that they contribute to a conversation a reference to the kind of which the previously mentioned individual is an instance. Hence, one is committed to discourse referents of different *types* in the representations: one for the particular instances and one for the general kind. Of course, this is not a challenge to the Novelty condition but a further confirmation that some version of it must be correct.[13]

[12] Lewis (2012, p. 318).

[13] An interesting twist of the data is the appositional phrases:

(1) a. A student1 walked into Sue's office and asked her about his exam.
b. Finally, a student2, one3 who walked into Sue's office and asked her about his exam, needed her help!

Two things to note about the use of appositives. First, suppose the second occurrence of "a student"—a student2—does trigger a general, category-like discourse referent, the appositional phrase—a student3—seems to be about something more specific, i.e., it is about a particular instance of a certain kind. One idea is that in its summary use, an indefinite is to be read into an appositive structure such that the general, purely existential meaning and the identity meaning are both captured.

One may wonder if, without the presence of the discourse particle "finally," the appositional phrase alone suffices the indefinite be read in the "summary" sense. That is,

(2) a. A student1 walked into Sue's office and asked her about his exam.

Further problems for the pragmatic approach concern the introductory uses and the deviance thereof that result from embedding indefinites in, for example, negations:

(17) a. Bill did not see a woman. # She was walking her dog.[14]
b. Bill did not see a woman who was walking her dog.

As "a woman" is embedded in negation in both (17a) and (17b), the pragmatic account should predict no addition of a file card, and a fortiori no later anaphoric expression on the presumably nonexistent object under discussion. Nevertheless, anaphora is permitted in (17b).

Consider also the contrast between (18) and (19):

(18) a. A woman walked in.
b. She looked gloomy.
(19) a. It is not the case that not every woman did not walk in.
b. # She looked gloomy.

Since Lewis equates content to truth conditions, (18a) and (19a) have exactly the same existential entailing content. But the use of anaphora in (19b) apparently is infelicitous. What is truth-conditionally equivalent to an indefinite (e.g., "a F is *G*" and "not every F is not *G*") does not possess the matching Licensing capacity. This discrepancy cannot be explained away by claiming that a file card is only introduced via linguistic acts that contain an explicit device of existential quantification. That response begs the question; it is neither an argument nor an explanation of the phenomena, but merely a restatement of the view that indefinites, but not other truth-conditionally equivalent expressions, provide the optimal, most perspicuous way of signaling a new object under discussion.

In addition, it is not transparent what answers the pragmatic account can supply regarding the ensuing minimal pair:[15]

b. A student2, one3 who walked into Sue's office and asked her about his exam, needed her help!

I think the answer is negative. It seems to me that somehow the general meaning of "a student" just is not available in discourse (2).

On the other hand, swapping the two noun phrases in the second sentence results in significant change of meaning:

(3) a. A student1 walked into Sue's office and asked her about his exam.
b. Finally, one3 who walked into Sue's office and asked her about his exam, a student2, needed her help!

Discourse (3) sounds much worse, if intelligible at all, and a purely existential, general meaning of a student2 is missing.

While these phenomena raise an interesting question of the ways "finally" interacts with noun phrases (as they are placed in different parts of the speech) that affect the availability of "summary reading," it goes far beyond the scope of this chapter to present a full theory of all the relevant details.

[14] The intended reading here is not one where "a woman" receives a wide-scope, de re interpretation.

[15] Such phenomenon is referred to as modal subordination in the literature. See, for example, Roberts (1987, 1989), Frank and Kamp (1997), and Asher and Pogodalla (2010).

(20) a. A wolf might come in.
b. It would will eat you first.

(21) a. A wolf might come in.
b. # It will eat you first.

The kind of analysis that Lewis advocates faces a general difficulty. The introductory uses of indefinites are supposedly the default, but subsequent use of anaphora is not ubiquitous. Various particles in the discourse can give rise to a control effect—negation and modals, for example, often are barriers to back referencing the object previously mentioned. Such control effects and the lack thereof, however, cannot be sufficiently justified by pragmatics. By contrast, a semantic account along the lines of dynamic theories offers a straightforward and more plausible explanation. The difference in the availability of subsequent anaphoric reference is analyzed in terms of a well-defined accessibility constraint, with no need to appeal to any equivocal notion of plan recognition.

4 Concluding Remarks

Let me conclude with some general morals from the foregoing discussion.

First, whether one takes the dynamic or the static stance, the right analysis must make recourse to a two-stage process. The dynamic theories have the two-level mechanism built in its very nature. The Novelty condition associated with indefinites applies at the level of the construction of the representation, that is, an indefinite invariably adds a new file card to the representation of the discourse; it is a separate issue if more than one file cards are mapped to the same object in the model. In this sense, there is nothing new or controversial about cases involving the summary uses of indefinites. What those examples do show, however, is that there are further constraints on the verification or satisfaction conditions of the dynamic discourse representations. Without a doubt, conversation participants are remarkably sensitive to how various linguistic expressions—e.g., discourse particles or other constructions that single rhetoric relations and discourse structures—affect the question under discussion and the at-issueness of a conversation; this is an issue that has drawn increasing attention in recent developments in the dynamic tradition.[16]

In contrast, the prospect of a natural and plausible two-step machinery is less rosy for the static approach that Lewis tries to defend and advocate. At the very least, she needs to justify the systematicity of Novelty and explain its occasional absence. Crucially, however, Lewis fails to make a strong case given the many difficulties that I manifest in the previous sections.

Second, one sees that what Lewis calls local plans get recognized regardless of the speaker's intention or plan for the entire discourse. This strikes me as evidence that local plans are nothing but the semantic content. Given the close tie between

[16] See, for example, Asher and Lascarides (2003) and Beaver et al. (2010).

plan and intention, a case in point is Bratman's (1984) distinction between "the intention to *A*" and "intentionally *A*." According to Bratman, when an agent intentionally *A*, she intends something, but she may not specifically intend to *A*. *A* is an intended action when it is an agent's intention to carry out *A*; by contrast, when one *A* intentionally, *A* may be an unintended consequence, or side effect, of one's intended action.

In the case of linguistic communication, a speaker's use of an indefinite is indicative of some communicative intention, but whatever that is, it need not be identical to the addition of a file card. In other words, one should distinguish between "introducing a file card intentionally" and "the intention to introduce a file card." Since a file card may be introduced by the use of an indefinite whether or not the speaker has a plan for the very introduction, the mechanism of file-card addition must operate in a way that is independent of planning and plan recognition. The best explanation is that the so-called local plan associated with a speaker's use of an expression is simply its semantic content, which is why they are at all recognizable. Uttering an indefinite triggers the introduction of a new file card; what the speaker plans or whether she has a plan is beside the point.

Lastly, there is a sense in which "how one ought to draw the distinction between semantics and pragmatics" is orthogonal to the central debate. What a theory is labeled as, be it dynamic, static, semantic, or pragmatic, is not what really matters. What really matters is how well the aid analysis handles the relevant linguistic phenomena, and whether in doing so illuminates our understanding of the underlying apparatus. While consistency, coherence, and comprehensiveness might to some degree be a judgment call, the data examined so far unmistakably support the approach that takes the dynamics of meaning very seriously.

Acknowledgments The chapter benefits greatly from discussions with Josh Dever, Ray Buchanan, and Hans Kamp. I am also thankful for the very helpful comments from Kiki Linton Wang, as well as feedback from the audience at the IEAS conference on Language and Action. The usual disclaimer applies.

References

Anette, F., & Kamp. H. (1997). On context dependence in modal constructions. In A. Lawson (Ed.), *Proceedings of SALT VII* (pp. 151–168). Ithaca: Cornell University.

Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge: Cambridge University Press.

Asher, N., & Pogodalla, S. (2010). A Montagovian treatment of modal subordination. In N. Li & D. Lutz (Eds.), *Proceedings of semantics and linguistic theory* (*SALT*) *20* (pp. 387–405). Ithaca: CLC Publications.

Beaver, D. I., Roberts, C., Simons, M., & Tonhauser, J. (2010). What projects and why. In N. Li & D. Lutz (Eds.), *Proceedings of semantics and linguistic theory* (*SALT*) *20* (pp. 309–327). Ithaca: CLC Publications.

Bratman, M. (1984). Two faces of intention. *The Philosophical Review, 93*(3), 375–405.

Groenendijk, J., & M. Stokhof. (1991). Dynamic predicate logic. *Linguistics and Philosophy, 14,* 39–100.

Gundel, J., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69*(2), 274–307.

Heim, I. (1982). *The semantics of definite and indefinite noun phrases.* Doctoral dissertation, University of Massachusetts at Amherst.

Heim, I. (1983). File change semantics and the familiarity theory of definiteness. In P. Portner & B. H. Partee (Eds.), *Formal semantics: The essential readings* (pp. 164–189). Berlin: De Gruyter.

Kamp, H. (1981). A theory of truth and semantic representation. In J. Groenendijk et al. (Eds.), *Formal methods in the study of language* (pp. 277–322). Amsterdam: Mathematics Center.

Kamp, H., & Reyle, U. (1993). *From discourse to logic*. Dordrecht: Kluwer Academic Publishers.

Lewis, K. (2012). Discourse dynamics, pragmatics, and indefinites. *Philosophical Studies, 158,* 313–342.

Pollack, M. E. (1992). The use of plans. *Artificial Intelligence, 57*(1), 43–68.

Roberts, C. (1987). *Modal subordination, anaphora, and distributivity.* Doctoral dissertation, University of Massachusetts at Amherst.

Roberts, C. (1989). Modal subordination and pronominal anaphora in discourse. *Linguistics and Philosophy, 12,* 683–721.

Russell, B. (1905). On denoting. *Mind, 14,* 479–493.

A Defense of the Knowledge Account of Assertion: From a Model-Theoretic Perspective

Syraya Chin-mu Yang

Abstract The main burden of this chapter is to defend the knowledge account of the norm of assertion—one must: assert *p* only if one knows *p, from a model-theoretical perspective*. I start with a classification of three different kinds of norms of assertion, i.e., semantic, pragmatic, and epistemic and focus on the epistemic normativity for assertion in this chapter. I take it as a starting point that the ultimate concern of making an assertion is to transmit true information via the embedded assertoric force in assertion that the agent performs, which can be in turn characterized by virtue of the agent's epistemic attitude toward the propositional content of whatever the agent asserts. Then, I examine three popular accounts of epistemic norms of assertion, including the truth account, the justified belief (or warrant) account, and the knowledge account.

I next examine, from a model-theoretic perspective, how to stipulate the required semantic rule for modal-operator A for assertion in the standard Kripke models for epistemic logic, containing the modal operators K (*knowing*), A (*asserting*), B (*believing*), and B^j (*having justified belief*), in accordance with each account and show that each has certain difficulties. I then show that the required semantic rule for Aφ can hardly be stipulated merely by the truth of φ in related accessible states, nor the appeal to Bφ or B^j. However, I show that the knowledge account can be defended by proposing a kind of models referred to as TWA-models ('TW' for Timothy Williamson). This is an application of my previous work, "TW-models for logic of knowledge-cum-belief" (Yang 2013). TW-models for a logic of knowledge and belief, a kind of Kripke models in character, with reflexivity as the sole accessibility relation, satisfy the main theses of Timothy Williamson's *knowledge first epistemology*. The required semantic rule for Kφ is stipulated in the standard way, but the semantic rule for B (correspondingly, for B^j) is given in a way such that the truth value of Bφ at a given state will be determined by the truth value of Kφ (correspondingly, Bφ) in some related states.

The construction of TWA-models will be specified, and the semantic rules for A will be given in a way such that the truth value of Aφ at a given state can be determined by virtue of the truth values of the corresponding Kφ in the accessible states. This will provide a model-theoretical justification of the knowledge rule. Some by-products will be mentioned briefly.

S. C. Yang (✉)
Department of Philosophy, National Taiwan University, Taipei, Taiwan
e-mail: cmyang@ntu.edu.tw

Tzu-Wei Hung (ed.), *Communicative Action,* DOI 10.1007/978-981-4585-84-2_3,

1 Assertion and Norms of Assertion

In ordinary discourse, assertion is a kind of intentional speech act,[1] performed by an agent (a speaker in character) by uttering a declarative sentence of the language in use[2] to claim, with a certain cognitive attitude,[3] explicitly or implicitly, that something holds. Granted that assertion is intentional action of some kind, it is a natural inclination to impose a certain type of norm on the agent so that a speech act of this kind can be distinguished from some other kinds, such as promising, commanding, etc. In particular, by virtue of the imposed norm, if there is any, the kind of substantial assertoric force will be embedded in an utterance of this sort, which will in turn not only express explicitly, but also guarantee that the agent does hold such a cognitive attitude toward the propositional content of what she asserts. It is in this sense that the proposed norm of assertion can be treated as the constitutive rule of assertion. In view of the specific significant consequence, during the last few decades a variety of norms of assertion have been proposed, but so far an agreement remains to be inaccessible.

At the moment, the proposed norms of assertion can be roughly classified into three different kinds. First of all, it has been taken for granted that in making an assertion, what is claimed to hold can be treated as the propositional content of the given sentence uttered, which is in turn taken as the meaning of the given sentence. From a linguistic point of view, it is widely agreed that words/expressions/sentences of a natural language in use by and large have so-called literal meaning. It is then disputable whether the agent should stick to the alleged literal meaning of words/expressions/sentences of the language in use when she utters a certain sentence in making an assertion. That is to say, would a certain kind of normative force be imposed on the agent so that she would stick to the meaning of the words/expressions/sentences in use in making assertion? Let us call this requirement of the norm of assertion the semantic normativity. A group of philosophers and linguists, typically the proponents of building-block theory and use theory of meaning, argue for the indispensability of semantic normativity; while the proponents of Quine's principle of the indeterminacy of translation (2013, Chap. 2) and Davidson's no language thesis (2005, p. 107) would show no sympathy for the alleged semantic normativity. It seems to me that whether semantic normativity should be imposed on assertion hinges upon a satisfactory theory of meaning. Unfortunately, so far this remains to be an open issue, and further investigation is beyond the scope of this chapter.

[1] See, for example, Williamson (2000, pp. 240, 241): "Assertion is a kind of speech act that we perform." Also Kemp (2013, p. 107): "assertion is individualistic in the sense that the matter of whether or not an agent asserts a proposition the agent expresses is entirely determined by the agent's intentions."

[2] See, for example, Williamson (2000, p. 258): "the default use of declarative sentences is to make assertions."

[3] See Mark (2010, pp. 163–164): "In order for an utterance to have assertoric force, it must also be subject to *the cognitive* and *social safeguards* that distinguish assertion…both from other illocutionary acts and from other forms of information transfer."

A second approach to the norm of assertion focuses on the pragmatic perspective, including the agent's intention and some other psychological aspects. Normative force of this kind can be called *pragmatic normativity*. A typical example would be Grice's Cooperative principle and the Maxims of Conversation. According to Grice (1989, pp. 26–28), the cooperative principle is a kind of norm governing all cooperative interactions among human beings and the conversational maxims can be treated as precisification of the cooperative principles that deal specifically with communication. However, so far there is no comprehensible and complete set of maxims to govern all kinds of assertion. Things get worse if we consider that just like breach of rules happens in all kinds of games all the time, in ordinary discourse, breach of the pragmatic norms of all kinds of speech acts happens quite often. It would be rather difficult to specify what the desired norm of assertion should be or should look like. It is well observed that the complexity of human behavior, and a fortiori, of human psychological activity, would make the appeal to a specification of the agent's intention problematic. It is beyond the scope of philosophical discussion at this stage to put forth a unifying pragmatic norm for assertion, in general, in ordinary discourse. I should therefore leave this kind of norm untouched as well.

In search of the norm of assertion a much more promising approach is to focus on the agent's cognitive attitude. It would be appealing to impose a certain kind of norm on the agent's assertion so as to display explicitly the agent's cognitive attitude toward what one asserts. If there is such a norm, the agent's cognitive attitude will be embedded in the uttered sentence simultaneously and we can treat this as the desired assertoric force. In other words, the norm of this sort, if there is any, will guarantee that the agent shows explicitly that she holds a certain *cognitive attitude* (such as "affirms" or "knows" or "believe," etc.) to what she claims to hold, that is, to ascribe a certain epistemic attitude to the propositional content of the sentence she uttered. For the sake of simplicity, let us call the assertoric force of this kind the epistemic normativity. In fact, speech acts are a kind of intentional activity embedded in a form of life, especially in ordinary discourse, appropriate for rational agents. Assertion should therefore display, explicitly or implicitly, the rationality of agents involved. A simple way to show that assertion would function in this way is to show that the agent would hold explicitly a certain epistemic attitude toward what she asserts. That is to say, by imposing an (*epistemic*) *norm of assertion*, certain *epistemic requirements will be imposed on appropriate (or correct) assertions*. As a matter of fact, by imposing such a norm, the agent can assert something only when she holds a certain epistemic attitude to what she intends to claim. This in turn shows the epistemic status of the propositional content of the very assertion. It is in this sense that the norm of assertion can be treated as a *constitutive rule* to govern the agents' speech acts of this kind. I shall, therefore, focus on the epistemic normativity of assertion in this chapter.

During the last few decades, at least three different accounts with regard to the required epistemic normativity of assertion have been proposed, including:

1. The truth account—one must: assert *p* only if *p* is true.
2. The justified belief account (or the warrant norm)—one must: assert *p* only if one has justified belief about *p*.
3. The knowledge account: "One must: assert *p* only if one knows that *p*."

Let us start with the truth account. Historically, the truth norm of assertion has its contemporary origin in Frege's work. According to Frege, asserting a proposition is to claim that it is true. Frege firstly assumed that there are thoughts which can be expressed in terms of sentences of a language in use. Moreover, thoughts can be grasped by the agent, and then judged as true or false. Frege (1984) emphasized that we must distinguish among (a) the grasp of a thought—thinking (or recognizing the propositional content of the sentence in use); (b) the acknowledgment of the truth of a thought—judgment (i.e., judging that the propositional content of the sentence in question is true); and (c) the manifestation of this judgment—assertion (explicitly expressing that the sentence is true.) On the basis of this distinction, Frege (1979, p. 2) maintained that

> The goal of scientific endeavour is truth. Inwardly to recognize something as true is to make a judgment, and to give expression to this judgment is to make an assertion (my italic).

In short, an assertion is an outward sign of a judgment. Apparently, Frege had already observed that there is a two-folded function hidden by the syntactic structure of assertoric sentences: (a) to express a propositional content; and (b) to express the truth of this content.

A significant consequence of this account is the factiveness of assertion, that is, asserting something is to ensure that it is true. Bearing in mind the factiveness of assertion in this sense, as Price (1998) points out, both the belief norm and the warrant (namely justified belief) norm are weaker than the truth norm. By contrast, the truth norm appears to be a bit weaker than the knowledge norm. Weiner (2005) argues in favor of the truth norm by claiming that a theory on which proper assertions must be true explains the data better than a theory on which proper assertions must be known to be true. After all, it is hard to characterize a general norm that governs all assertions to the extent that the agent must know whatever she asserts. It is possible to explain the cases that motivate the knowledge account by postulating a general norm that *assertion would be true*, combined with conversational norms (typically, Gricean maxims) that govern all kinds of speech acts in general. As a matter of fact, Frege seemed to notice this so he noted that no sign or other gesture, not even the predicate "is true" could be sufficient for assertoric force (Frege 1979, p. 251). That is, some more things are required for assertion apart from the truth of the propositional content asserted. The trouble is that there is hardly a way to specify exactly what the extra requirement is. Moreover, if we appeal to so-called "would be true," instead of "being true," it would be hard to retain the factiveness of assertion.

Apart from this, there are several arguments against the truth account of assertion. Koethe (2009, p. 632) claims that if the assertibility conditions for *p* coincide with those for the claim *to know that p*, and if these in turn are merely identified with the truth conditions for *p* and for the claim *to know that p*, it would follow that the truth conditions for *p* and for the claim *to know that p* are the same, which has the absurd consequence that one knows everything that is true. Recently, Pelling (2011) also showed that a liar sentence alike assertion, viz., when the agent asserts that "this assertion is improper," would constitute a counterexample to the truth account of assertion. Apparently, if the agent's assertion is true then the very assertion is improper; if it is false, then the assertion is proper. Still, for some who accept the nonfactiveness of assertion, truth account is too strong. For example, Hinzen (2013, p. 130) even claims that we could have some practice of assertion and communication but no concept of truth.

We have already noted that when an agent asserts something by uttering a sentence, say *s*, what the agent shows is not only the propositional content of the very sentence *s* (presumably, say the proposition *p*), but also shows some assertoric force, which can be embedded in the agent's *epistemic states* toward what is asserted. Viewing the acquired normativity from this aspect, the truth account would have nothing to do with the agent's epistemic states. In fact, the truth norm cannot even show that assertion is a kind of intentional speech act. Moreover, saying that a sentence *s* is true adds no assertoric force to the propositional content of *s*. To impose a certain assertoric force to an utterance, the agent must explicitly show a certain epistemic state toward the propositional content of what one is uttering.

Following this line of thought, a better account of the norm of assertion is to appeal to belief, for instance, Hindriks (2007) and Oppy (2013). At the first glance, it seems attractive to appeal to belief. At any rate, sometimes we may want to assert something for which we have no idea whether it holds or does not hold. However, taking belief as the norm for assertion appears too weak to be accepted. After all, as Williamson (2000, p. 255) points out, false beliefs are often reasonable in a commonplace. Perhaps, a better way is to appeal to justified belief: When an agent asserts something, at least she must have some good reason or evidence or justification to support, or to warrant, whatever she asserts.[4]

However, what exactly the very notion of justification is supposed to mean? How to assess in what sense and to what extent a belief is justified? It is to be noted that there are at least two different senses in which a belief is said to be justified. The first one means the justification of the propositional content of what the agent believes; so it is required that there is a justification of the propositional content of what is believed. The second one means that the agent's believing something, say φ, is well-justified.

Sticking to the first sense of justified belief, Kvanvig (2009) claims that:

(RBNA) One may assert *p* only if one *has a reasonable belief* that *p*.

According to Kvanvig, (RBNA) can explain the data that Williamson took to support his knowledge account, i.e., the impropriety of asserting lottery propositions and Moorean assertions. (RBNA) also provides a unified treatment to a number of counterexamples to knowledge account, such as the issue concerning prediction. Douven

[4] For the sake of simplicity, we may for the time being treat "One rationally believes…," "One reasonably believes…," "One has justification to believe…," "One justified believes…," etc., as synonyms, similarly for the phrases "justified belief," "reasonable belief," "rational belief," "having justification for belief," "warranted belief." Henceforth, I shall use the term "justified belief" or justification.

(2006) proposes a similar approach, the so-called rational credibility account.[5] But it would be hard to characterize what a reasonable belief is, or should be.

At present a much more promising approach is to appeal to justification logic. Roughly speaking, justification logic is a formal framework which incorporates epistemic assertion that there is a justification t for the propositional content of a certain formula (i.e., what the agent knows/believes), in symbols "t:φ," into some sort of modal logic, typically some kind of epistemic logic for knowledge/belief.[6]

From a logical point of view, justification logic is well-established. A group of distinct systems (axiomatization), with appropriate semantics has been proposed. The completeness theorem can be proved as well. Nevertheless, from a philosophical point of view, there are a number of misgivings over the appeal to justification logic, as far as the appropriateness of the philosophical concept of justified belief is concerned. First of all, it is well-observed that in justification logic, the format "t:φ" is by and large taken as primitive, in that no direct analysis of what it means for a justification operator t to justify a formula φ has been offered. In fact, justification logic is merely intended to characterize the very relation (between a collection of justifications and formulas) in terms of some sort of axiomatization. Things would get worse, as far as axiomatization is concerned. For it has been shown that almost every major modal logic, such as K, T, K4, S4, K45, S5, and some others, has an exact justification logic counterpart, e.g., J_0, J, JT, J4, J45, etc. (Note that this is by no means to claim that any modal logic has a reasonable justification logic counterpart, typically, the logic of provability GL.) We then have a family of justification logic based on a variety of choice of operations. Now the question is: Which one is the best or the most appropriate one for a philosophical interpretation of justification involved? In particular, when some justification logic accepts factiveness of justification, some reject. Moreover, the ontological status of whatever the modal operator for justification refers to is rather problematic. We should take for granted that justifications are abstract entities which have structure and operations on them. This would make the desired models for assertion much more complicated.

Of course, in the traditional analysis of knowledge in terms of justified true belief, a justification for the propositional content of a given sentence is in general to certify knowledge in the sense that based upon the given justification the agent can announce that she knows what is justified. But, should we accept this analysis

[5] According to Douven, the rational credibility account is implied by two of our basic commitments, namely (a) our aiming to be rational and (b) the belief-assertion parallel, according to which belief is subvocalized assertion. It is noteworthy that Douven criticized that Williamson is committed to the belief-assertion parallel, because he holds that "occurrently believing *p* stands to asserting *p* as the inner stands to the outer" (Williamson 2000, p. 255). But this criticism seems to ignore Williamson's account of belief in his knowledge first epistemology that belief can only be characterized in terms of knowledge.

[6] The recent development of justification logic can be found in a series of work by Artemov (1995, 2001, 2006, 2008), Artemov and Nogina (2005), and Fitting (2005). The initial justification logic system, namely, the Logic of Proofs (LP), was firstly introduced in Artemov (1995); a more general approach to common knowledge based on justified knowledge can be found in Artemov (2006). Fitting (2005) firstly provides epistemic semantics and established completeness for LP.

of knowledge? In particular, Williamson and his friends would stick to his "E = K" Thesis—only knowledge can serve as evidence. Following this line of thought, justified belief, understood in this sense would either fail to hold or collapse into knowledge account.

A second sense of justified belief would pay attention to the justification of the agent's believing something as a whole. Lackey (2007) argues that perhaps we should give up the seemingly much more subjective conception of justification proposed by Kvanvig or Douven which can be expressed by the phrase such as "The agent reasonably believes that – – –." Instead, we should stick to the following form:

$$\text{(RTBA)}\;\; \text{One must assert } p \text{ only if it is reasonable to believe } p.$$

What is required for this account is not a justification of what is believed (i.e., the propositional content of *p*) but a justification of the agent's act of believing itself. It seems to me that to make sense of the notion of "it is reasonable to believe…" we need to presuppose so-called common knowledge/belief, an open issue so far. Be that as it may, the appeal to justified belief can hardly be factive. So later, Kvanvig (2011) adopts a stronger version of justification, knowledge-strength justification in character, called *epistemic justification*:

$$\text{(RBNA}^{*}\text{)}\;\; \text{One must : assert } p \text{ only if one has justification to believe that } p.$$

Unfortunately, it is hard to characterize the alleged knowledge-strength epistemic justification, unless we have already had a clear picture about the role the concept of knowledge plays in the desired account of assertion. This would lend this account collapse into knowledge account.

So far, it seems that the most appealing one is the knowledge account.[7] Recall that on Frege's truth account of assertion, making an assertion presupposes having a judgment; and having a judgment presupposes the grasp of a thought in turn. Now if we stick to the epistemic normativity of assertion, it seems perfectly reasonable to claim that without grasping the propositional content of a sentence *p* and then making a judgment, we cannot assert that *p*.[8] Then the knowledge account follows automatically!

There are several good reasons for the knowledge account. Williamson has appealed to two arguments. Firstly, he argues that "Only knowledge warrants asser-

[7] For example, Unger (1975, Chap. 6), Williamson (1996, 2000, Chap. 11), DeRose (2002, p. 180): "One is positioned well-enough to assert that P iff one knows that P"; Stanley (2005, pp. 10, 11): "[A]ssertion is…conceptually connected to knowledge…*one ought only to assert what one knows*"; Hawthorne (2004, p. 23): "The practice of assertion is constituted by the rule/requirement that one assert something only if one knows it"; to mention a few. Perhaps, the earliest version of knowledge account can be found in G. E. Moore's work, when he claims that "by asserting *p* positively you *imply*, though you don't assert, that you know that *p*" (1962, p. 277, see also 1960, p. 125).

[8] Typically, this is the root of the well-known *Moore's paradox*—"*P*, and I don't know that *P*" (Moore 1962).

tion" (2000, p. 243), based on his well-known E = K thesis—Only knowledge can serve as evidence, and hence only knowledge can justify knowledge. Moreover, for Williamson, we express and communicate our knowledge by making assertion (Williamson 2000, p. 238). The skeptics have raised several doubts on the knowledge account. Weiner (2005) offers two arguments against this knowledge account. Firstly, there is no general norm that the speaker must know whatever she asserts. By contrast, there are cases in which it can be entirely proper to assert something, such as predictions and retrodictions, which are generally acceptable in the absence of knowledge for the agent. Secondly, the truth account, together with Gricean mechanisms, thus can explain the data that motivated the knowledge account. For our assertions to be appropriate (or correct), not only must they be true, we must have reason to believe them true. The hearer thus is entitled to conclude that the speaker has some warrant for her assertion. In many cases, the most likely warrant combined with truth will be enough for knowledge. When it is obvious that the speaker's warrant would not be enough for knowledge, assertion without knowledge is permissible (Weiner 2005, p. 238). At the very beginning of this chapter, I have noted the distinction among epistemic normativity and pragmatic normativity of assertion. It then strikes me that Weiner's criticism seems to stick to the category of pragmatic normativity. Some others, such as Lackey (2007, p. 596), argue that knowledge account is "too strong a requirement of assertion," and defend the justified (or reasonable) belief account. Koethe (2009) argues against knowledge account on the ground that it would not accept the (epistemic) possibility of asserting something false.

It is not my intention here to dwell on the pros and cons for any of these accounts of the norm of assertion. The main burden of this chapter is to defend the knowledge account, from a model-theoretical point of view. The underlying thought is to show that if an account of epistemic norm of assertion works well, the semantic treatment of assertion should be able to be characterized in terms of the semantic treatment of the corresponding epistemic formulas involved in the proposed norm. That is, we should be able to stipulate the required semantic rule for assertions by some other modal operator, as the proposed account suggests.

In what follows I shall firstly present the standard models, a kind of Kripke models in character, for epistemic logic in general. Next, I add to the original language in use one extra modal operator "A" for assertion. Then, I briefly examine whether the required semantic rule for "A" can be stipulated based on the proposed account of norm of assertion. That is to say, all that I want to do is to see if a formal semantic rule for the epistemic operator for assertion "A" can be explicitly characterized in terms of the proposed epistemic requirements. It is in this sense we can reasonably say that the proposed norm of assertion would pave a way to the stipulation of the required semantic rule for the assertion-operator so that the truth value of an assertion, say $A\varphi$, in a given state will be determined based on the truth value of the sentence of the specified corresponding epistemic attitude in some related states. This simply echoes Williamson's claim that the required norm is *constitutive rule of assertion*, and is not derived from any general rules governing conversation.

2 Epistemic Norms of Assertion: From a Model-Theoretic Perspective

To examine the appropriateness of an account of the norm of assertion from a model-theoretic perspective, we need a setting, viz., the standard Kripke models for epistemic logic. Let us fix a standard language for epistemic logic, say L_K: $p|\neg\varphi|\varphi\rightarrow\psi|\mathrm{K}\varphi$. (For a logic of belief, replace Kφ by Bφ; for a logic of knowledge-cum-belief, add Bφ to L_K , for a logic of assertion, substitute Aφ for Kφ…, and so on.)

A standard Kripke model M for epistemic logic in general M is a complex of the form $\langle S, \sigma, R\rangle$, where
S: a nonempty set of states, or possible worlds,
σ: $S \rightarrow (P \rightarrow \{T,F\})$, an assignment of a truth value of $\{T, F\}$ to the propositional letters of the language in use in every state,

$R \subseteq S \times S$ (the required accessibility relation among all states).

The semantic rules for p, $\neg\varphi$ and $\varphi\rightarrow\psi$ are standard; while the rule for Kφ is stipulated as

$$M,w \vDash \mathrm{K}\varphi \quad \text{iff for all states } u \text{ in } S \text{ with } Rwu\, M,u \vDash \varphi.$$

It is somewhat interesting to notice that all the modal operators K, B, and A are universal in that it will range over all states accessible from the given one. The difference among the semantic treatments of these three operators merely lies in the accessibility constraint imposed. At the moment, a variety of accessibility relations have been proposed. It has been observed that the constraint of the imposed accessibility relation on a model will validate certain corresponding characteristic formula(s) including:

- $(K)\,\Box(\varphi\rightarrow\psi)\rightarrow(\Box\varphi\rightarrow\Box\psi)$, (The Distributive Law)
- $(T)\,\Box\varphi\rightarrow\varphi$, (The Truth Axiom; Factiveness)
- $(4)\,\Box\varphi\rightarrow\Box\Box\varphi$, (Positive Introspective Axiom)
- $(5)\,\neg\Box\varphi\rightarrow\Box\neg\Box\varphi$, (Negative Introspective Axiom)
- $(D)\,\Box j$®$\phi\Box\phi j\,[\neg\Box(\varphi\wedge\neg\varphi)$, or $\neg\Box\bot]$.

Note that here we may replace □ by K to get an axiom concerning knowledge, the same goes for belief (or assertion), i.e., replacing K by B (A, respectively). Here we shall assume the well-known correspondence between a variety of accessibility relations and the aforementioned characteristic formulas.[9]

[9] For instance,
1. R to be reflexive, i.e., $\forall w\in S$, Rww, the Kripke models would satisfy (T).
2. R to be transitive, i.e., $\forall w, u,v\in S$, $Rwu\wedge Ruv\rightarrow Rwv$, the Kripke models would satisfy (4).

Now, let us start with the truth account. The standard way to set the required semantic rule for Aφ is this,

$$M, w \vDash \mathrm{A}\varphi \quad \text{iff for all states } u \text{ in } S \text{ with } Rwu \quad M, u \vDash \varphi.$$

Admittedly, such a standard semantic treatment appears to be in favor of the truth account. However, from a model-theoretic perspective, the truth account of assertion has some problems. First of all, as we have just noted, on the standard Kripke models for epistemic logic, all operators K, B, and A are universal in that Kφ/Bφ/Aφ is true in a state *s* only if φ is true in all states accessible from *s*. The difference at the issue merely lies in the proposed accessibility relation, especially the acceptance of reflexivity for K; while a rejection of reflexivity for B. Now, if we accept the truth account, we must accept the standard semantic rule for A: Aφ is true in a state *s* only if φ is true in all states accessible from *s*, with a further specification of the conditions for the required accessibility relation. But what kind of condition is required for assertion? To my knowledge, no such required accessibility relation has been proposed as yet,[10] and I admit that it is beyond my comprehension to propose one, either. Moreover, assertion is a kind of intentional speech act with the agent's epistemic attitude toward the propositional content of whatever she asserts. Clearly, the truth account shows nothing related to these aspects. These considerations make the truth account unacceptable.

We now move to the justified belief account. At present, a great number of logics of belief has been proposed and we do have corresponding semantic rule for modal operator B:

$$M,\ s \vDash \mathrm{B}\varphi \quad \text{iff for all states } t \text{ in } S \text{ with } Rst\ M, t \vDash \varphi$$

But I know of no semantic treatment for justified belief account of assertion. (I have already discussed the approach appeal to the justification logic.) Still, notice that

3. *R* to be equivalence, i.e., with reflexivity, transitivity, and also symmetry, the Kripke models would satisfy (*T*)+(*4*)+(*5*).
4. *R* to be transitive and Euclidean ($\forall w, u, v \in S$, $Rwu \wedge Rwv \rightarrow Ruv$), but not reflexive, the Kripke models would satisfy (*K*)+(*4*)+(*5*).
5. *R* to be serial, i.e., $\forall x \in S \exists y Rxy$ (*no end-point included*), the Kripke models would satisfy (*D*).

Again, taking CPC—classical propositional calculus as the underlying system, we may have some well-known modal systems, such as System **K** (=CPC+(*N*)+(*K*)), System **T** (=**K**+(*T*)), System **S4** (=**T**+(*4*)), System **S5** (=**S4**+(*5*)). Of course, the rule of Necessitation is required—(*N*) Form $\vdash\varphi$ one can get $\vdash\Box\varphi$.

Usually we may have systems **K, T, S4, S5** for knowledge and assertion. But some may reject (*T*), so the modal operator A would behave like the modal operator for believing. Accordingly, it is suggested that we should take so-called **KD45** or *weak* **S5**, by adding (D_B) to **K45** yields the so-called **KD45** or *weak* **S5**. Some take both **K45** and **KD45** as appropriate for a logic of assertion.

[10] A quite popular one is to appeal to **S5**. But in this case, there would be no difference between knowledge and assertion. *Assertion collapses into knowledge*.

neither belief nor justified belief is factive. This would threat the epistemic value of assertion.

So now, what remains is only knowledge account. If this account works well, we should have a semantic rule for A:

$$M, s \vDash \mathrm{A}\varphi \quad \text{iff for all states } t \text{ in } S \text{ with } Rst\ M, t \vDash \mathrm{K}\varphi$$

A great advantage of this semantic treatment is that we are no longer to be bothered by the issue concerning what kind of a second accessibility relation for assertion should be imposed. Moreover, it also suggests that assertion is factive. This appears to be much more congenial to Frege's original thought.

However, from a model-theoretical point of view, there are several misgivings over the knowledge account. Firstly, it is striking that a model-theoretic treatment to the knowledge account of the normativity of assertion must presuppose a certain epistemic logic of knowledge as the underlying system. But at the moment there is a variety of epistemic logic systems for knowledge and no decisive conclusion showing which one is the best. Even Williamson offers no appropriate semantic treatment of assertion in his knowledge-first epistemology.

Secondly, as is widely observed, the logic of knowledge, or epistemic logic in general, has been suffering from the problem of logical omniscience, a problem which can be found in normal modal systems in general. It is well-known that a modal operator M in a modal system is called *normal* if the system contains *Necessitation*—(*N*) $\vdash\varphi \Rightarrow \vdash \mathrm{M}\varphi$, as a rule of inference and (*K*)—$\mathrm{M}(\varphi\to\psi)\to(\mathrm{M}\varphi\to\mathrm{M}\psi)$, as an axiom. Correspondingly, a modal system is *normal* if its primitive modal operators are normal, such as *T*, *S4*, *S5* for knowledge, and *S4*, *K45*, and *KD45* for belief. In a normal epistemic logic, unrestricted applications of (*N*) and (*K*) would render a truism that if the agent knows/believes a proposition φ, and φ logically implies ψ, then the agent knows/believes ψ as well. As Levesque (1984, p. 198) puts it, "at any given point (world), the set of sentences considered to be believed is closed under logical consequence." Also, the agent knows/believes all valid sentences. However, even for a rational agent in ordinary discourse, as a resource-limited being, she may know/believe φ and $\varphi\to\psi$, but does not know/believe ψ due to a failure of drawing any connection between φ and ψ. It is patent that the logic of assertion is by and large a *normal modal system* as well, as the logic contains (*N*) $\vdash\varphi \Rightarrow \vdash\mathrm{A}\varphi$ and (*K*)—$\mathrm{A}(\varphi\to\psi)\to(\mathrm{A}\varphi\to\mathrm{A}\psi)$. Now, if we accept the knowledge account of assertion, we should accept a semantic consequence that we are entitled to assert anything that the agent knows, which sounds odd.

Thirdly, the knowledge account has to deal with, or at least to explain away the so-called Fitch's paradox—a knowledge variant of Moore's paradox, say

"*p* and I don't know *p*,"

in symbols, "$p \wedge\neg\mathrm{K}p$." Paradoxes of this kind seem unavoidable if we stick to the knowledge account. Granted the knowledge account of assertion, we should take it for granted that $\mathrm{A}p\to\mathrm{K}p$. So when the agent asserts (1), i.e., $\mathrm{A}(p\wedge\neg\mathrm{K}p)$, $\mathrm{K}(p\wedge\neg\mathrm{K}p)$ follows immediately. Then a simple derivation would render a contradiction, as we do have both $\mathrm{K}p$ and $\neg\mathrm{K}p$. It is striking that any proposed treatment for knowledge account of assertion must be able to explain away paradox of this sort.

The remaining part of this chapter will then be devoted to show that it is possible to construct an appropriate semantics for a logic of assertion which is based on the knowledge account. Moreover, the aforementioned problems can be dealt with.

3 TW-Models for a Logic of Knowledge and Belief

In Yang (2013), I proposed a kind of models, called TW-models, for an epistemic logic of knowledge and belief, which satisfy the main theses of Williamson's *knowledge first epistemology*, proposed in his *Knowledge and its Limits*.[11] Here I can only give a brief description of TW-models without detailed explanation.

Fixed a propositional language for an epistemic logic with modal operators "K" for knowledge and "B" for belief, say **L**:: $p| \neg\varphi| \varphi\rightarrow\psi|K\varphi|B\varphi$. A TW-model, as a complex, can be then described in what follows:

$$M = \langle S, \sigma, \mathrm{R}, \ \delta\rangle,$$

S : A non-empty set of states, or possible worlds;

σ: $S\rightarrow(P\rightarrow\{T, F\})$, an assignment of a truth value of $\{T, F\}$ to the propositional letters of the language in use in every state.
$R \subseteq S\times S$: a partial ordering with reflexivity to serve as the required accessibility relation among all states.
δ: $S\rightarrow\wp(\mathbf{L})$, such that for any $s\in S$, $\delta(s)\subseteq\{\varphi|\ M, s\vDash\varphi, \varphi\in\mathbf{L}\}$, in particular, for any state $u\in S$ if $\forall t\in SRut\rightarrow t=u$, $\delta(u)=\{\varphi|\ M, u\vDash\varphi, \varphi\in\mathbf{L}\}$, i.e., $\varphi\in\delta(u)$ iff M, $u\vDash\varphi$.

The semantic rules for atomic formulae, negation, and material implication are standard. We only consider semantic rules for Kφ and Bφ, respectively, in what follows:

$$(\mathrm{K}^S)\ M, s \vDash \mathrm{K}\varphi \quad \text{iff } \forall t \in S(Rst \rightarrow M, t \vDash \varphi) \wedge \varphi \in \delta(s).$$

[11] I have summarized there Williamson's knowledge-first epistemology in terms of the following theses:

- Knowing is a state of mind.
- Knowing is factive.
- The broadness of knowing (Externalist approach).
- The primeness of knowing (Knowledge first!).
- Take knowledge as central to our understanding of belief.
- Cognitive-homeless thesis.
- The knowledge account of evidence—One's knowledge is just one's evidence.
- The knowledge account of assertion—Assert *p* only if one knows that *p*.

I acknowledged that TW-models would not deal with the knowledge account of assertion, and this will be settled in TWA-models in this chapter.

$$(\mathrm{B}^{S})\ M, s \vDash \mathrm{B}\varphi \text{ iff}$$
$$(\mathrm{i}) M, s \vDash \mathrm{K}\varphi;\ \text{or}$$
$$(\mathrm{ii}) \exists u \in S\{[Rsu \wedge \forall t \in S(Rut \rightarrow u = t)] \wedge M, u \vDash \mathrm{K}\varphi\},$$

that is, there is a self-isolated state u in S accessible from s, such that Kφ is true in *u* (and a fortiori, φ is true in *u*).

The introduction of δ, referred to as *ipk*-function, is to signify Williamson's original notion of the agent's *being in a position to know* a proposition in a state. For Williamson, merely true in all nearby cases would not be sufficient for an agent to know it. It may happen that in a given state, some propositions appear to be true in all nearby cases but the agent is not in a position to know them, so that the agent may not know them. But, I have noted that Williamson's original idea is somewhat confusing; I thereby proposed a refined notion of "*being actually in a position to know*" instead. Formula $\varphi \in \delta(s)$ will be interpreted as saying that the agent is *actually* in a position to know φ in *s*.

One can see clearly that the first condition in (K^{S})—"$\forall t \in S(Rst \rightarrow M, t \vDash \varphi$"—simply follows Hintikka's proposal that knowledge can be viewed as "*truth throughout the logical space of possibilities that the agent considers relevant*," or in Williamson's (2000, p. 17) words, "we know *p* only if *p* is true in nearby cases." The second condition in (K^{S})—"$\varphi \in \delta(s)$"—indicates the requirement that to know φ, the agent must *be actually in a position to know* φ in the given state. After all, *being true in all relevant states* is not sufficient for knowledge; in order to know what actually is, the agent must be *actually* in a certain condition so that when the case obtains, the agent can *actually* get what obtains.[12]

It is noteworthy that in the introduction of the *ipk*-function, we put forth a special case "for any state $u \in S$ if $\forall t \in SRut \rightarrow t = u$." This intends to introduce a special kind of state, referred to as *self-isolated states* in the sense that a self-isolated state can never access to some other states apart from itself, although it is accessible from some other states. The condition "$\varphi \in \delta(s)$ iff $M, s \vDash \varphi$" intends to stipulate that in a self-isolated state, the agent is actually in a position to know what happens, which in turn implies that whatever is true should be taken as a piece of knowledge. That is, $M, u \vDash \varphi$ iff $\varphi \in \delta(u)$ iff $M, u \vDash \mathrm{K}\varphi$. The proposed self-isolated states will play a crucial role in the stipulation of semantic rule for Bφ.

The first condition, in fact a disjunct, in (B^{S})—$M, s \vDash \mathrm{B}\varphi$ iff $M, s \vDash \mathrm{K}\varphi$—merely shows the rationality of human agent: One must believe whatever one knows. This also indicates that TW-models satisfy (KB) $\mathrm{K}\varphi \rightarrow \mathrm{B}\varphi$. While the second condition, a second disjunct, indicates that the agent may believe a proposition φ in *s* simply on the ground that φ is true in some self-isolated state accessible from *s*. Note that the specification of the *ipk*-function in self-isolated states—$\delta(u) = \{\varphi |\ M, u \vDash \varphi\}$—together with the reflexivity accessibility relation already shows that $M, u \vDash \varphi$ iff

[12] Accordingly, we need to add into the language in use one more modal operator, say I_{K}, so that $\mathrm{I}_{\mathrm{K}}\varphi$ is to mean "the agent is actually in a position to know φ." The corresponding semantic rule for I_{K} will be stipulated in the following way: $M, s \vDash \mathrm{I}_{\mathrm{K}}\ \varphi$ iff $\varphi \in \delta(s)$.

$\varphi \in \delta(u)$ iff M, $u \vDash K\varphi$. I argued that this formulation can capture Williamson's (2000, pp. 46–47) heuristic account of belief in terms of knowledge: To believe *p* is *to treat p as if she knows p*, that is*, to treat p in ways similar to the ways in which subjects treat propositions which they know*.[13]

One can see two noticeable characteristics of the construction of TW-models, which make it different from the standard Kripke models for epistemic logic. The first one concerns the constraint on the required accessibility relation on models. As we have noted, in standard Kripke models two distinct accessibility relations are required for modal operator K and B, respectively. However, in TW-models all that is required is reflexivity as the sole accessibility relation. The second one has something to do with the semantic rule for belief operator: The semantic rule for Bφ can be stipulated in the way that the truth value of Bφ at a given state can be determined by virtue of the truth values of the corresponding Kφ in the accessible states, though the semantic rule for Kφ is given in the standard way.

I have also noted two by-products of the proposed TW-models. Firstly, an extra epistemic modal operator for justified beliefs, namely, $B^j\varphi$ (*A rationally/reasonably believes* φ) and the semantic rules for B^j can be stipulated in the way that the truth value of $B^j\varphi$ at a state will be determined by virtue of truth values of the corresponding Bφ in some related state, which can be determined in turn by the semantic value of Kφ in some related states. Secondly, the problem of logical omniscience can be avoided on TW-models. It can be shown further that Williamson's underlying thought that there is no good reason to accept a belief-based account of assertion can be justified.

It is then rather appealing to apply TW-models to an epistemic logic containing knowledge operator and assertion operator. For, if we follow this line of thought and if knowledge account of assertion is acceptable, we should be able to construct a kind of models wherein the rule for Aφ can be stipulated in the way that the truth value of Aφ at a given state can be determined by virtue of the truth values of the corresponding Kφ in the accessible states. Hence, we need not put forth any further constraint on the required accessibility relation. This is what the remaining part of this chapter is devoted to.

4 TWA-Models for Knowledge Account of Assertion

We now turn our attention to the construction of desired models for the proposed knowledge account of assertion, an extension of TW-models in essence, referred to as TWA-models. The aim is to show that the semantic rule for Aφ can be stipulated in the way that the truth value of Aφ at a given state can be determined by virtue of the truth values of the corresponding Kφ in the accessible states. Also the problem of omniscience can be avoided in the proposed TWA-models.

[13] Williamson (2000, p. 207) notes that "if evidence is what justifies belief, then knowledge is what justifies belief." Also, "knowledge, and only knowledge, justified belief" (2000, p. 185).

Again, fix a language in use, say $\mathbf{L}_{\mathbf{A}} :: p \mid \neg\varphi \mid \varphi \to \psi \mid \mathrm{K}\varphi \mid \mathrm{B}\varphi \mid \mathrm{A}\varphi$. A TWA model is substantially an extension of a TW-model with an extra function.

$$\lambda : S \to \wp(\mathbf{L}_{\mathbf{A}}) \text{ such that for any } s \in S, \lambda(s) \in \delta(s).$$

Strictly speaking, the function λ intends to capture the idea that when an agent asserts something, taken as a kind of intentional speech act, she is doing this with an intention. The intended interpretation for a formula $\varphi \in \lambda(s)$ is to mean that *the agent A asserts φ with an intention.*[14] The condition "$\lambda(s) \subseteq \delta(s)$" shows that when the agent has an intention to assert φ, she must be actually in a position to know what she intends to assert. After all, assertion is a kind of intentional speech act, and it would be hard to accept that someone would assert something that she does not know. So it seems beyond reasonable doubt to set as a requirement that the agent must be actually in a position to know φ in s. Moreover, in view of the assertoric force of the knowledge account, it is striking that the agent must *know* that she *knows whatever she intends to assert*. Accordingly, we need a semantic rule for the modal operator for assertion A as what follows:

$$(\mathrm{A}^{S}) \qquad M, s \vDash \mathrm{A}\varphi \quad \text{iff} \quad \forall t \in S(Rst \to M, t \vDash \mathrm{K}\varphi) \wedge \varphi \in \lambda(s) \wedge \mathrm{K}\varphi \in \delta(s).$$

The first condition, $\forall t \in S(Rst \to \mathrm{M}, t \vDash \mathrm{K}\varphi)$, simply sticks to the knowledge account of assertion: "One asserts p only if one knows p, and in turns, only if $\mathrm{K}\varphi$ is true in all nearby cases." The second condition, $\varphi \in \lambda(s)$, indicates that *knowing* φ *in all relevant states* is not sufficient for assertion, to assert φ, the agent must *be with an intention to assert* φ in the given state. The third condition merely suggests that the agent must *be actually in a position to know* that she *knows whatever she intends to assert*. This has a quite significant consequence in that $\mathrm{A}\varphi \to \mathrm{KK}\varphi$ holds in TWA-models, though she may not know what she is doing, namely asserting something.[15]

Some further remarks are noteworthy. First of all, it can be shown that TWA-models can explain the failure of the justified belief account. I have also introduced an extra modal operator B^j, to signify the notion of justified belief, and proposed a semantic rule for it as what follows:

$$(\mathrm{B}^{jS})\ M, s \vDash \mathrm{B}^{j}\varphi \text{ if } \forall t \in SRstM, t \vDash \mathrm{B}\varphi, \text{ and } \mathrm{B}\varphi \in \delta(s)$$

The semantic rule simply says that the agent A has a justification of believing φ, only if A believes φ in all nearby cases (i.e., in all states accessible from the actual one), and she is actually in a position to know that she believes φ, a fortiori knows that she believes φ. This implies not only $\mathrm{B}^j\varphi \to \mathrm{B}\varphi$ but also $\mathrm{B}^j\varphi \to \mathrm{KB}\varphi$. Now, on the

[14] As Davidson (2001, p. 90) rightly remarked, there are no such conventions governing the formation of intentions. So I can only put forth a primitive function here.

[15] Davidson (2001, p. 91) notes that "It is a mistake to suppose that if an agent is doing something intentionally, he must know that he is doing it." This indicates that $\mathrm{A}\varphi \to \mathrm{KA}\varphi$ would not hold. But it seems beyond reasonable doubt to claim that the agent must know that she knows what she asserts, otherwise, it would be difficult to show how she could do this intentionally.

proposed knowledge account, if Aφ holds in s, then Kφ is true in all states accessible from s. Also, Bφ is true in all states accessible from s. Accordingly, B^jφ is true in s. This justified that Aφ →B^jφ.[16]

However, if we accept the warrant account, the semantic rule for A would be:

$$\left[\mathrm{A}(\mathrm{B}^j)^S\right]\ M,s \vDash \mathrm{A}\varphi \quad \text{iff} \quad \text{for all states } t \text{ in } S \text{ with } RstM,t \vDash \mathrm{B}^j\varphi$$

The problem concerning the factiveness of assertion remains. For, the notion of justified belief does not imply truth on TW-models. Moreover, the agent may not be able to know that she has justified belief already. By contrast, the proposed semantic rule (A^S) for Aφ shows that the agent not only knows whatever she intends to assert but also knows that she knows that. This would be much more close to the ordinary usage of assertion.

Secondly, the well-known formula (4) for assertion, i.e., Aφ→AAφ can be invalidated in TWA-models, though we do have Aφ→KKφ. Williamson (1995) has argued that (4_A), i.e., Aφ→AAφ, should not hold based on his anti-luminosity thesis. Davidson, as we have just mentioned, also claims that doing something intentionally does not imply that the agent knows what she is doing. So if Aφ→AAφ holds, given that Aφ→Kφ, we would have Aφ→KAφ, which is not acceptable.

Thirdly, the so-called Fitch's paradox—a knowledge variant of Moore's paradox can be dealt with on TWA-models. It is obvious that given A(p∧¬Kp) and Ap →Kp, K(p∧¬Kp) follows immediately. But then on TWA-models, we may not be able to derive Kp or K¬Kp from K(p∧¬Kp). This is because at a certain state s, though (p∧¬Kp)∈δ(s) so that K(p∧¬Kp) holds but it may not be the case that p∈δ(s) or ¬Kp∈δ(s) so that for the agent Kp or K¬Kp may not hold. Therefore, the Fitch paradox is then solved.

Finally, the problem of logical omniscience for assertion can be dealt with in a way similar to the way we deal with the problem of logical omniscience for knowledge. The introduction of the function λ will play the same role as the introduction of the function δ.

A final remark. Williamson (2000, pp. 260–262) considers two more norms of assertions. The first one is known as believing-knowledge account:

(*ABK*) One is in a position to assert that p just in case (only if) one believes that one knows that p.

Along this line of thought, van Benthem (2010, p. 148) maintains that the force of assertion in ordinary discourse appears to be stronger than belief but weaker than knowledge. Hence, (*ABK*) would be a better candidate for an analysis of assertion. It is clear that both Aφ→φ and Aφ→Bφ hold automatically because both Kφ→φ and Kφ→Bφ hold in TW-models. Moreover, given Kφ∈λ(s) and λ(s)⊆δ(s), it is obvious that if the agent asserts φ, then she knows that she knows φ. In short, Aφ→KKφ. A natural consequence of this is that we do have Aφ→BKφ (as Kφ→Bφ). But if the truth value of Aφ is to be characterized in terms of BKφ in all nearby cases, the

[16] In short, on TWA-models, the knowledge account of assertion satisfies the following conditions: (a) Aφ→φ; (b) Aφ→Bφ; (c) Aφ→BKφ; and (d) Aφ→B^jφ.

agent should know that she knows φ in all cases accessible from every nearby case. Semantically, this requirement is much stronger and more complicated than knowledge account. A similar argument with appropriate modification would be sufficient enough to reject the RBK-rule (i.e., one must: assert *p* only if one rationally believes that one knows *p*).

The second one says:

(ABB^{j}K) One must: assert that *p* just in case (only if) one believes that one has a rational belief that one knows that *p*.

I admit that it is beyond my comprehension to figure out the required conditions for Aφ to be characterized in terms of the truth value of BB^{j}Kφ in all nearby cases. After all, we have rejected the RBK-rule. And it would be more difficult to establish the required truth condition for Aφ in terms of BB^{j}Kφ.

I hope that based on the proposed TWA-models, we may be able to put forth an axiomatization of logic of assertion. But at least, TWA-models will be able to justify that knowledge account of assertion is the best candidate from a model-theoretic perspective.

Acknowledgment Previous versions of this paper were presented at *Asian Workshop on Philosophical Logic* (February 15-17, 2012, Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan), and *Workshop on Logic, Philosophy, and Computation* (October 19, 2012, Institute of Information Science, Academia Sinica, Taiwan). I am deeply grateful to all participants on these occasions for their helpful remarks and comments. I also thank anonymous referees and Duen-Min Deng for their valuable comments and suggestions.

References

Artemov, S. (1995). *Operational modal logic* (Technical Report MSI 95–29, Cornell University). http://www.cs.cornell.edu/Info/People/artemov/MSI95–29.ps.

Artemov, S. (2001). Explicit provability and constructive semantics. *Bulletin of Symbolic Logic, 7*(1), 1–36.

Artemov, S. (2006). Justified common knowledge. *Theoretical Computer Science, 357*(1/3), 4–22.

Artemov, S. (2008). The logic of justification. *Review of Symbolic Logic, 1*(4), 477–513.

Artemov, S., & Nogina, E. (2005). Introducing justification into epistemic logic. *Journal of Logic and Computation, 15*(6), 109–1073.

DeRose, K. (2002). Assertion, knowledge, and context. *Philosophical Review, 111*(2), 167–203.

Davidson, D. (2001). *Essays on actions and events (2nd ed.)*. Oxford: Clarendon Press.

Davidson, D. (2005). *Truth, language, and history*. Oxford: Clarendon Press.

Douven, I. (2006). Assertion, knowledge, and rational credibility. *Philosophical Review, 115*(4), 449–485.

Fitting, M. (2005). The logic of proofs, semantically. *Annals of Pure and Applied Logic, 132*(1), 1–25.

Frege, G. (1979). *Posthumous writings* (trans: P. Long & R. White). Oxford: Basil Blackwell.

Frege, G. (1984). Thought. In B. McGuinness (Ed.), *Gottlob Frege's collected papers-on mathematics, logic, and philosophy* (pp. 351–372). Oxford: Basil Blackwell. (Original work published 1918)

Grice, P. (1989). *Studies in the way of words*. Cambridge: Harvard University Press.

Hindriks, F. (2007). The status of the knowledge account of assertion. *Linguistics and Philosophy, 30*(3), 393–406.

Hinzen, W. (2013). Truth, assertion and the sentence. In D. Greimann & G. Siegwart (Eds.), *Truth and speech acts: Studies in the philosophy of language* (pp. 130–156). London: Routledge.

Hawthorne, J. (2004). *Knowledge and lotteries*. Oxford: Oxford University Press.

Kemp, G. (2013). Assertion as a practice. In D. Greimann & G. Siegwart (Eds.), *Truth and speech acts: Studies in the philosophy of language* (pp. 106–129). London: Routledge.

Koethe, J. (2009). Knowledge and the norms of assertion. *Australasian Journal of Philosophy, 87*(4), 625–638.

Kvanvig, J. (2009). Assertion, knowledge, and lotteries. In P. Greenough & D. Pritchard (Eds.), *Williamson on knowledge* (pp. 140–160). Oxford: Oxford University Press.

Kvanvig, J. (2011). Norms of assertion. In J. Brown & H. Cappelen (Eds.), *Assertion: New philosophical essays* (pp. 233–250). Oxford: Oxford University Press.

Lackey, J. (2007). Norms of assertion. *Noûs, 41*(4), 594–626.

Levesque, H. J. (1984). A logic of implicit and explicit belief. *Proceedings of the National Conference on Artificial Intelligence, August 6–10, 1984, University of Texas at Austin: AAAI-84*, 198–202.

Mark, J. (2010). *Assertion*. New York: Palgrave Macmillan.

Moore, G. E. (1960). *Ethics* (2nd ed.). Oxford: Oxford University Press. (Original work published 1912)

Moore, G. E. (1962). *Commonplace book: 1919–1953*. London: Allen & Unwin.

Oppy, G. (2013). Norms of assertion. In D. Greimann & G. Siegwart (Eds.), *Truth and speech acts: Studies in the philosophy of language* (pp. 226–249). London: Routledge.

Pelling, C. (2011). A self-referential paradox for the truth account of assertion. *Analysis, 71*(4), 688.

Price, H. (1998). Three norms of assertibility, or how the MOA became extinct. *Philosophical Perspectives, 12,* 241–254.

Quine, W. V. (2013). *Word and object*. Cambridge: MIT Press. (Original work published 1960)

Stanley, J. (2005). *Knowledge and practical interests*. Oxford: Oxford University Press.

Unger, P. (1975). *Ignorance: A case for scepticism*. Oxford: Oxford University Press.

van Benthem, J. (2010). *Modal logic for open minds*. Stanford: CSLI.

Weiner, M. (2005). Must we know what we say? *Philosophical Review, 114*(2), 227–251.

Williamson, T. (1995). Does assertibility satisfy the S4 axiom? *Crítica*: *Revista hispanoamericana de filosofía, 27*(81), 3–25.

Williamson, T. (1996). Knowing and asserting. *Philosophical Review, 105*(4), 489–523.

Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.

Yang, S. C.-M. (2013). TW-models for logic of knowledge-cum-belief. *The Proceedings of 12th Asian Logic Conference*. Singapore: World Scientific Publishing Co.

Part II
Action and Bodily Awareness

When Actions Feel Alien—an Explanatory Model

Timothy Lane

Abstract It is not necessarily the case that we ever have experiences of self, but human beings do regularly report instances for which self is experienced as absent. That is, there are times when body parts, mental states, or actions are felt to be alien. Here, I sketch an explanatory framework for explaining these alienation experiences, a framework that also attempts to explain the "mental glue" whereby self is bound to body, mind, or action. The framework is a multidimensional model that integrates personal and sub-personal components, psychological and neural processes. I then proceed to show how this model can be applied to explain the action-related passivity experiences of persons suffering from schizophrenia. I argue that a distinctive phenomenological mark of these experiences is that they are vividly felt, unlike ordinary actions (those taken to belong to self), and I seek to explain these heightened sensory experiences from within the proposed framework. I also propose hypotheses concerning such phenomena as thought insertion and anarchic hand syndrome that are motivated by this framework. Finally, I argue that the proposed model and view of self-experiences is consistent with several aspects of and theories of consciousness, especially theories which indicate that consciousness is more likely to be engaged when we are dealing with novelty or error—e.g., when self seems to have gone missing. I conclude by recommending that if we wish to learn about self, we would be well advised to attend closely to those times when it seems absent.

"The self is an absence…" (Sorensen 2007, p. 450)
"…we can feel the absence of the self…but not the presence of the self." (Prinz 2012a, p. 148)

1 Introduction

It is not obvious that we ever have experiences of self, nor is it obvious that there is any such entity that is properly regarded as a self. But human beings do regularly report instances for which self is experienced as absent. That is there are times when body parts,

T. Lane (✉)
Taipei Medical University, Institute of Humanities in Medicine and Shuang Ho Hospital, Brain and Consciousness Research Center
e-mail: timlane@tmu.edu.tw

Academia Sinica, Institute of European and American Studies

National Chengchi University, Research Center for Mind, Brain, and Learning

Tzu-Wei Hung (ed.), *Communicative Action,* DOI 10.1007/978-981-4585-84-2_4,

mental states, or actions seem to be alien. In somatoparaphrenia, an arm, despite being attached to the body, is experienced as belonging to someone else. In schizophrenia, a thought, despite only being reportable from the first-person perspective, seems not to belong, to have been inserted into the stream of consciousness. And, in anarchic limb, an arm performs actions that do not belong to self, actions that feel alien.

Collectively, these and numerous other phenomena suggest that Prinz and Sorensen, who are quoted in the epigraph, have identified a key characteristic of self, or at least an important issue concerning the nature of epistemic access to self. It is known more by its absence than by its presence. In ordinary circumstances, there may not be any self-experiences or "I-qualia" at all (cf. Prinz 2012b, p. 214). But when subject to an illusion or suffering from pathology, persons do sometimes have vivid experiences that self is absent.

The purpose of this essay is to extend a line of thought begun previously (Lane 2012): that is, these various alienation experiences can be explained from within a unified framework. The "mental glue" (Klein 2013a, p. 90) in virtue of which self is bound to its body, its mental states, and its actions is all or largely implicit. When this relationship is "unstuck" is when we experience the absence of self.[1] The thesis extended here is that absence experiences just are the *meta-awareness (MA) that components of experience are clustering in atypical ways, even though those experiences are being processed on a sub-personal level as highly self-related*. Here it is argued that action alienation is, at least in part, explainable in these terms.

2 Problems of Belonging

Previously I (cf. Gallagher 2012, pp. 207–211; Lane and Liang 2011) have argued that even when only one person is positioned to introspect[2] on or report a mental state, it should never be presupposed that the mental state *belongs to* that person. "Belonging," or what I have elsewhere referred to as "mental ownership" (Lane and Liang 2010), is always a contingent relationship. This is not to say that mental states or conscious experiences can fly about untethered, as though they were baseballs. Clearly, mental states are more like dents than baseballs, in that they cannot exist on their own. But to say that some organism must *host* them is one thing; to say that

[1] There is something potentially misleading about talk of a *relationship between self* and body parts, mental states, or actions. Talk of belonging or mental glue in such contexts suggests a two-place relation, which is not the case, at least not if one of the relata is taken to be an experience of self (cf. Prinz 2012b, p. 231). Although for expository purposes I do write in such a way that a relationship of this type might seem to be implied, as I explain below, "self" here is taken to be a distinctive type of neural activity, not an experience in and of itself. What Baars et al. (2003) and Baars (2007) metaphorically refer to as the brain's "observing self" is a useful way to think of this relationship.

[2] Throughout this manuscript I use "introspection" in what Shoemaker (1994, p. 258) dubs the "humdrum" sense, referring to information we have access to that is expressed in such remarks as "it itches," "I'm thirsty," and so forth.

they necessarily belong to the self who introspects upon and reports them is something else again (Klein 2013a, p. 90; Lane 2012, p. 260).

Cases of craniopagus twins can usefully illustrate the belonging relationship (e.g., Stone and Goodrich 2006). A very recent case involves two sisters who are connected at the head in a manner that forces them to face away from one another (Bor 2012, pp. 28–31). It seems that a neural bridge connects their thalami,[3] and that this connection makes it possible for them to have some degree of joint access to sensory processing: for example, one girl will introspectively sense the thirst of her sister and proceed to reach for a cup of water that she hands to her conjoined sibling. In effect, when such sharing occurs, each of the siblings is able to distinguish between those sensations "that *belong to* self and those that *belong to* her sister" (Bor 2012, p. 29).[4] In other words, one sibling might have taste, tactile, or visual sensations that belong exclusively to her. But the other sibling can become aware of these sensations via introspection. The latter sibling though, despite having this seemingly direct access to the sensations does not feel that they belong to her. The mental glue is absent.

Succinctly, problems of belonging (POBs) are those instances wherein from the third-person perspective a body part, a mental state, or an action would seem to belong to a given person, but from the first-person perspective the experience is of alienation. As mentioned previously, somatoparaphrenia, schizophrenic thought insertions, and anarchic limb are all instantiations of belonging's absence. But POBs can also occur in the reverse condition: that which from the third-person perspective could not possibly be said to belong is experienced as belonging. When the rubber hand illusion is successfully induced, a detached, artificial hand is experienced by the participant as belonging to self (e.g., Lane et al. 2013). When thought and behavior are manipulated in certain ways, as in "I Spy" experiments (Wegner 2002, pp. 74–78),[5] actions performed by someone else are felt to have been performed by self. And in cases of synesthesia for pain, the usual self–other distinction that enables us to experience empathy without literally experiencing what is felt by others, collapses, such that what we observe in others triggers pain that belongs to self (Fitzgibbon et al. 2010). In short, POBs do not occur in quotidian circumstances; instead, they are associated with pathology, illusion, or other atypical phenomena.

Lane (2012) and Klein (2013a, b) have described a variety of cases of different types, wherein $n=1$, that illustrate POBs. These include, for example, instances of visual sensations and episodic memory. In both cases, the person who is uniquely situated to report on these experiences reports that they seem not to belong to self. For the case of vision, visual states are not immediately taken as belonging to self (Zahn et al. 2008). There is a time lag between initial awareness and subsequent relating of the images to self. In the case of episodic memory, memories that sat-

[3] The thalamus projects a large number of axons to all parts of the cortex; the cortex projects an even greater number of axons to the thalamus (Jones 2007).

[4] Italics not contained in the original.

[5] These experiments attempt to capture, under controlled conditions, certain aspects of what persons experience when they play with an Ouija board.

isfy all the conditions for counting as episodic are felt not to belong to self (Klein and Nichols 2012), despite the fact that episodic memories are typically taken to be uniquely self-involving. Although the total number of recorded cases is small, it might be that they are underreported, because they are so counter-intuitive, and because natural language was not designed to express such aberrancies (Lane 2012, p. 259; Klein 2013a, p. 101, b, p. 11). In the interest of both focusing attention on these phenomena, and in order to introduce POBs as they are related to action, an explanatory model is adumbrated in the next section.

3 An Explanatory Model of Belonging

If we take cases for which, from the third-person perspective, there is no doubt but that mental states are realized in virtue of a person's brain (along with endocrine and immune systems, as well as spinal cord) activity, but that the person describes experiencing those states as alien or estranged, as belonging to someone else, pain asymbolia can serve as paradigmatic of POBs. Typically when pain states are experienced, sensory–discriminative and affective–motivational aspects are conjoined (Auvray et al. 2010). In other words, not only is a person able to identify such discriminate features of pain as its qualitative character (e.g., cramping, shooting, stabbing, or burning), location, duration, and intensity, those features are intimately—what might seem to be intrinsically—associated with suffering and aversive reactions. In the normal course of events that is, the sensory–discriminative and the affective–motivational are experienced as a whole, faithfully reflecting what we mean when we say "I am in pain"—it hurts and we desire relief.

But in cases of pain asymbolia, the sensory–discriminative and the affective–motivational dissociate (Grahek 2007, pp. 51–71), a dissociation that is often associated with lesions to the parietal operculum and the posterior insula. Here, although persons so afflicted are able to report pain sensations, they also report being unbothered. That they are not bothered is confirmed by the absence of withdrawal behaviors, an easy willingness to cooperate with pain testing, and the inability to learn appropriate avoidance behaviors, despite normal perception of both deep and superficial pain (e.g., Berthier et al. 1988, pp. 42–43). In such persons pain is, so to speak, shorn of its painfulness.

Pain shorn of its painfulness can occur because our pain system bifurcates: the path that engages the affective–motivational projects to the limbic system, while the path that enables fine sensory discrimination projects to the cortex. In some cases, the effects of such disconnection can be so profound that patients do more than report not being bothered, not experiencing painfulness. Some patients report acute awareness of sensory–discriminative aspects of pain, as well as the feeling that those introspectively accessed pain sensations "seem to belong to someone else, not to me" (Sierra 2009, p. 150).[6]

[6] Somewhat similar reactions to pain have been reported for those who have undergone surgical ablation of pathways linking the thalamus with parts of the frontal lobes (Klein 2013a, pp. 91–93).

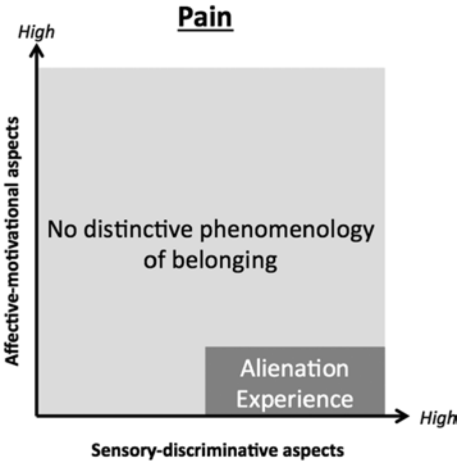
Pain
High
Affective-motivational aspects
No distinctive phenomenology of belonging
Alienation Experience
High
Sensory-discriminative aspects

Fig. 1 Principle of confounded expectations: the case of pain asymbolia (Reprinted with permission of Springer Science+Business Media, from Lane, T. (2012). Toward an explanatory framework for mental ownership. Phenomenology and the Cognitive Sciences, 11, 251–286)

Figure 1 delineates this relationship. Under normal conditions, those tacitly assumed, whenever sensory–discriminative aspects of pain are vividly felt, so too are the affective–motivational aspects. To abstract away from this particular case, we could say that this phenomenon is an instantiation of a general principle, what I refer to as mental state clustering (MSC). That is to say, mental states are not experienced in isolation, but that they cluster is something we do not notice until something goes amiss.[7]

Here the affective–motivational aspect of pain has gone missing, as is indicated by the *y*-axis. This confounding of tacit expectations aptly illustrates what I refer to as the principle of confounded expectations (PCEs). One interpretive option available to the person whose expectations are confounded in this way is that the pain states do exist, but that they do not belong to self.

Such characterization of pain should not surprise, because the neural substrates for self-pain and other-pain are distinct (Ochsner et al. 2008, p. 153). One region in particular is worth noting: a large portion of the mid-insula that is *posterior to* the region activated in common for both self- and other-pain (Ochsner et al. 2008, p. 153) evinces distinctively higher levels of activation for self-pain. Since pain asymbolia is often associated with lesions to the posterior insula, from the first-person perspective it might be a perfectly reasonable inference to attribute the pain sensation to "someone else."

At least one comprehensive theory of the brain seems broadly consistent with the notion of PCE. In its essence, the idea is that brains are hypothesis-testing machines dedicated to minimizing the error of their predictions about sensory input

[7] Instances of mental states failing to cluster in tacitly expected ways are plentiful. To cite just one other example, those who suffer from motion blindness have otherwise normal visual experiences of the external world, but their ability to perceive motion is greatly impaired (e.g., Zihl et al. 1983). Typically visual perceptions of color and shape cluster with motion. But in these rare cases, clustering fails.

received from the world (Clark et al. 2013; Friston 2009; Hohwy 2013) or from the body (Seth et al. 2012). If the brain is just such a machine, prior probabilities would lead it to predict standard clustering. But when such predictions or expectations are repeatedly confounded by sensory inputs, in order to minimize prediction error, the brain would generate novel hypotheses. In this case, the pains belong to someone else. Although this would not be literally true, at least not in the sense that the pains were being realized in someone else's body, assuming that the posterior insula lesions remain as they are, there might not be sensory inputs of the right sort to bring about further adjustment in the brain's hypothesis.

But even if brains are hypothesis-testing machines, MSC and PCE could not be sufficient to explain POBs. To illustrate with one respect in which these will need to be augmented, consider the difference between pain asymbolia and Capgras syndrome. In the former, patients are aware of pains that seem not to belong to self; in the latter, patients are aware of familiar faces that seem oddly alien (Bortolotti 2010, pp. 68–73). As is the case with pain asymbolia, Capgras syndrome is sometimes explained as a confounding of expectations, something having been left out. As with pain asymbolia, what is left out is expected, affective response (Ellis and Lewis 2001).

Although both syndromes involve PCE, only pain asymbolia is relevant to POB. Capgras syndrome concerns not what belongs to self, but what is *familiar* to self. Accordingly, an explanatory framework adequate to handle POBs requires some means of distinguishing that which belongs to me from that which is familiar to me. Fortunately, recent discoveries concerning self-related processing (SRP) at the neural, sub-personal level enable the making of such a distinction.

For more than a decade, Northoff (e.g., 2013a, p. 79; 2013b, pp. 255–256) has been documenting ways in which the brain's cortical midline structures (CMSs)[8] and the subcortical midline structures (Northoff and Panksepp 2008) are involved in determining whether perceptual stimuli are self-related. Employing fMRI, EEG, and MRS, he and his research team have identified certain regions of interest as well as patterns of electrical and biochemical activity that enable us to distinguish between those stimuli that are and those that are not self-related. Of special relevance here are the results of a meta-analysis (see Fig. 2), which show that at the neural level we distinguish among those stimuli that are *self*-related (e.g., one's own name or face), those that are *familiar* (e.g., names or faces of close family members), and those that are neither, the *other* condition (e.g., names or faces of well-known, but not familiar, people). Several regions seem to be implicated in SRP, and much less so in the familiar or other conditions (Northoff 2013a, pp. 258–259; Northoff and Qin 2011). Especially noteworthy in this regard is the perigenual anterior cingulate cortex.

How then might MSC, PCE, and SRP be combined within an explanatory framework, so as to enable distinguishing phenomena related to belonging from

[8] CMS regions include the perigenual anterior cingulate cortex, the dorsomedial prefrontal cortex, and the posterior cingulated cortex. The CMS overlaps with Feinberg's (2009, pp. 152–155) "integrative self-system" and with the default-mode network (Raichle 2010), which has also been implicated in self-referential processing.

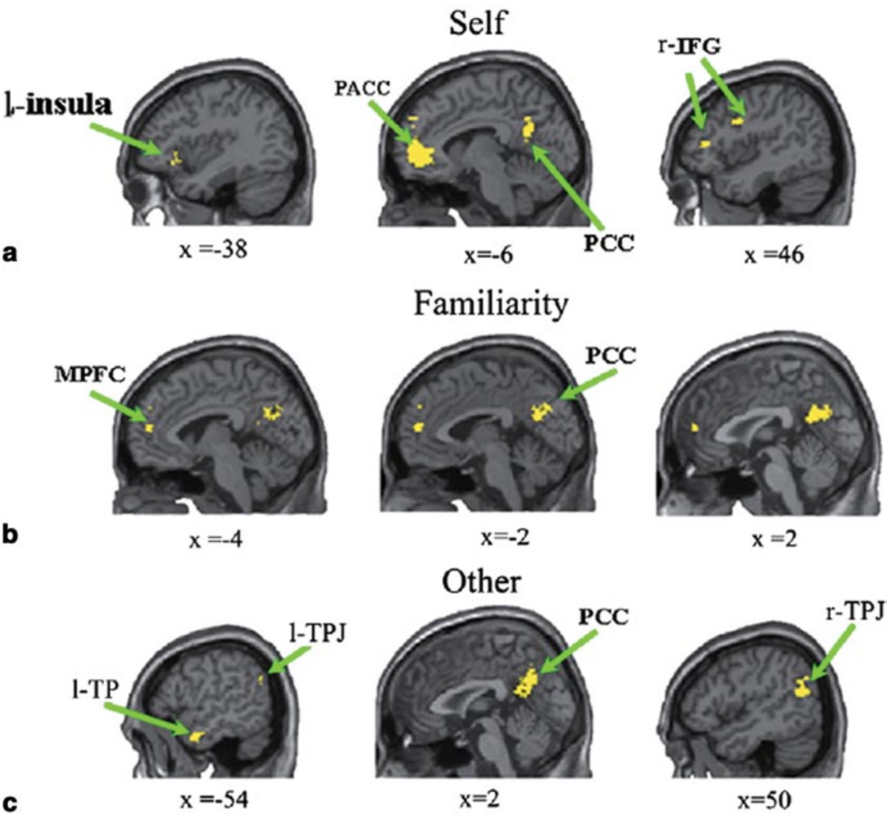

Fig. 2 The activated clusters for three conditions: self (**a**), familiarity (**b**), and other (**c**), based on a multilevel kernel density analysis. *MPFC* medial prefrontal cortex, *PACC* perigenual anterior cingulate cortex, *PCC* posterior cingulate cortex, *l-TP* left temporal pole, *l-insula* left insula, *l-TPJ* left temporoparietal junction, *r-TPJ* right temporoparietal junction, *r-IFG* right inferior frontal gyrus (Reprinted with permission of Elsevier, from Qin, P., Northoff, G. (2011). How is our self related to midline regions and the default-mode network? NeuroImage, 57(3), 1221–1233)

phenomena related to familiarity or to other conditions? The confounding of expectations can be illustrated by a simple 2-D model, as with Fig. 1 previously. But in order to distinguish among different types of confounded expectations, we need to add a third dimension, as is depicted in Fig. 3.

Indeed, the framework need not be restricted to explaining self or familiar experiences. It can even be extended to include other phenomena as, for example, de-realization, the feeling that one is cut off from the outside world, that it seems "unreal" (Sierra 2009, pp. 38, 39). Articulate patients frequently ascribe the feeling of "unreality" to the absence of affective coloring. They describe the world as seeming distant, flat, or artificial. If the loss of affect, or its distortion, is a significant component of de-realization, then the model proposed here seems capable of explaining a spectrum of phenomena that includes self, familiar, and other experiences.

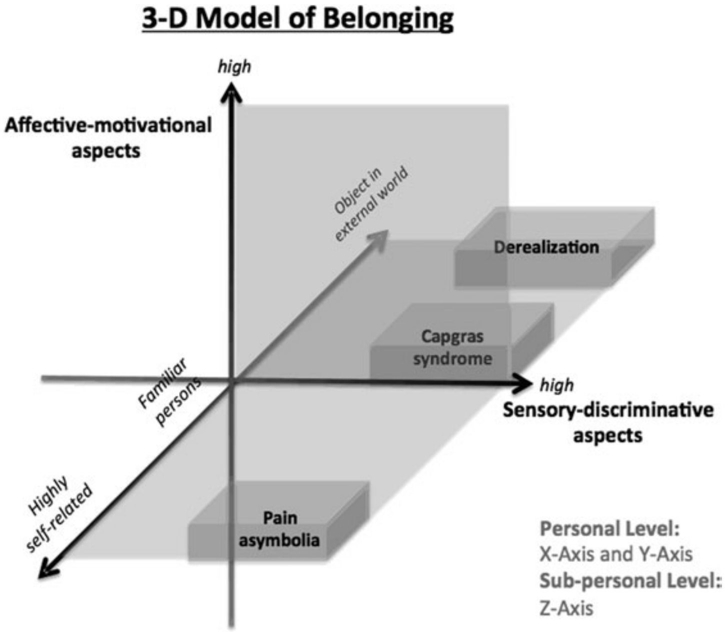

Fig. 3 *z*-axis added to model to illustrate role of sub-personal, self-related processing

This framework is not just a post hoc reconstruction. It suggests lines of inquiry when new instances of POB are encountered. For example, recall the case of episodic memory reported by Klein and Nichols (2012). In this instance, the memories satisfy all conditions for counting as episodic, yet they are felt not to belong to self. A prediction motivated by this model is that if we encounter another case of this type, we should investigate to determine whether the patient is experiencing loss of affect, perhaps caused by lesions similar to those observed in cases of pain asymbolia, while SRP remains intact.

Some philosophers (e.g., McDowell 1994) have expressed reservations about mixing personal and sub-personal elements within the same explanatory framework. The personal level of description involves, inter alia, description of conscious mental states and the subjects who experience those states; the sub-personal level, on the other hand, involves the mechanistic explanations of the objective sciences, including explanations of the sort invoked here, those that make reference to brain regions and patterns of electrical and biochemical activity (cf. Davies 2000; de Pinedo-Garcia and Noble 2008). But because my concern in proposing this framework is to suggest testable hypotheses (cf. Crick and Koch 2003, p. 119) and to make predictions concerning heretofore unexamined phenomena (cf. Hempel 1965, p. 365), what matters is whether or not the framework does indeed prove fruitful.

What is more, empirical discoveries that connect sub-personal properties to the personal level can diminish worries about seemingly mongrel approaches to explanation (Shea 2013). In the matter at hand, the relevant empirical discoveries involve judgments as regards whether or not a stimulus (visual, auditory, etc.) is self-related. A growing body of evidence suggests that these judgments can be predicted based upon either pre-stimulus neural activity or post-stimulus neural activity that occurs before subjects have conscious experience of the stimuli (e.g., Northoff 2013a; Northhoff et al. 2014). In other words, empirical studies suggest that a point of contact between sub-personal properties—neural activity in anterior CMSs—and the personal level—judgments concerning self-relatedness—has been identified.

Still, even after incorporating the *z*-axis, two problems remain. First, the *x*- and *y*-axes are only one example of confounded expectations. In the case of pain asymbolia, "pains" that do not hurt are experienced, contrary to what is normally expected of pain. But expectations can be confounded in multiple ways, even if we assume that the experiences are highly self-related. So here the problem does not just concern distinguishing among self, familiar, and other. There is a need to distinguish among the different ways in which mental states can cluster.

This point will be developed in more detail in the next section, when action is treated in some detail. But to illustrate briefly another way in which expectations can be confounded, consider the case described by Zahn et al. (2008) that was mentioned previously. Here, the person who reports visual states describes them as not immediately experienced as belonging to self. Instead, there is a time lag between initial awareness and subsequent relating of the images to self, and this description of the patient's phenomenology seems to be at odds with standard definitions of phenomenal consciousness. Carruthers (2000, p. 14), for example, defines phenomenally conscious events as "ones that we recognize in ourselves, non-inferentially, or 'straight off,' in virtue of the ways in which they feel to us, or the ways in which they present themselves to us subjectively." Similarly, Rosenthal (2002, pp. 408–422, 2005, pp. 343, 344) describes conscious mental states as feeling "direct," "unmediated," or "noninferential."

The case of visual sensations delayed seems to be an instance of phenomenally conscious events that confound expectations concerning how mental states of this type should be experienced. That is they seem not to be straight off, direct, or unmediated. These characterizations of normal conscious experience do not necessarily imply that there should be no time lag, but they suggest as much. And other examples from the empirical literature also suggest that time is a critical factor, that we expect immediacy.

Sass and Parnas (2003, p. 438) describe a patient suffering from schizophrenia whose experiences are illustrative. The patient "reported that his feeling of his experience *as his own experience* only 'appeared a split-second delayed'." Although the pathologies are distinct, the mode of describing the experience of conscious experience is similar. The implication I derive from these cases and the standard philosophical characterizations of conscious experience is that the latter are informed by quotidian cases, those upon which our expectations are based. In other words, philosophical intuitions accurately capture normal human expectations. But these aberrant cases show that immediacy is not a necessary condition for conscious experience, and that when there is a time delay a POB is reported.

The characterization of POBs thus far given and the implications of PCE, however, suggest two respects in which the model must be elaborated. First, since expectations can be confounded in different ways, it must be the case that the *x*- and *y*-dimensions can have different contents. That is, awareness of different types of atypical MSC can yield POBs. Therefore, a more comprehensive model must allow for such variability among the *x*- and *y*-dimensions.

Moreover, atypical clustering of mental states accompanied by sub-personal determinations of the degree of self-relatedness could not be sufficient to explain POBs. For POBs to occur, expectations must be confounded, expectations *about how lower level mental states* should cluster. In other words, to take the case of pain asymbolia as an example, not only must there be an awareness of the sensory–discriminate contents of pain, there must be an MA that something—the affective–motivational aspects—are missing. The same would be true for instances of temporal lag.

Meta-awareness, or meta-consciousness, is a process whereby individuals "take explicit note of the current contents of consciousness" (cf. Schooler et al. 2011, pp. 319, 321). In cases of POB, one becomes meta-aware that the contents of consciousness are clustering in atypical ways. It should be noted that this incorporation of meta-mental states is not the same as meta-cognition that is employed in higher-order thought (HOT) theories of consciousness (e.g., Lau and Rosenthal 2011; Rosenthal 2005). The difference is that in HOT theories, the lower level states are not presumed to be conscious in the absence of the meta-level; the meta-level is a necessary condition for conscious experience to obtain. Indeed, according to HOT theories, it is even the case that lower-level states are unnecessary (Lane and Liang 2008). Here, however, MA is becoming aware of lower-level mental states that are not dependent upon the meta-level in order for them to be conscious experiences.

The complete model can then be depicted as in Fig. 4. Here, in this 4-D model, the *x*- and *y*-axes of the original 3-D model are collapsed into one dimension, the MSC. This adjustment provides three benefits: one, it allows for an unlimited variety of distinct types of mental clustering. Two, it allows for the possibility that clustering differs by degree, ranging from the routine to the bizarre. And, three, it allows for the likelihood that clustering should not be thought of merely as a 2-D phenomenon. The sensory-discriminative and affective-motivational example depicted in Fig. 1, as well as the example of action alienation to be presented below, might seem to suggest that for MSC one need only consider two dimensions. But there is no empirical reason to believe that two dimensions are adequate to cover all instances of POB.

As for the sub-personal level, here identified as SR, nothing is changed. As is the case with the 3-D model, it indicates the degree to which a sensory input is assessed as self-related. MA is new to this version of the model. Here too, difference is by degree, and it might be just as much the result of high-level as lower-level activity. The example of alienation described by Sass and Parnas (2003) that was presented previously can illustrate the greater involvement of higher level activity. They explain the delay in the psychotic patient's report of a delay in feeling experiences to be his own as due

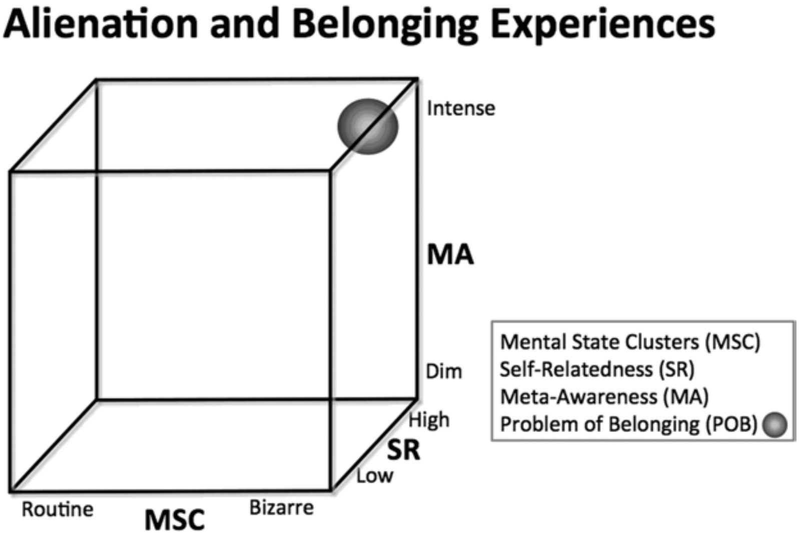
Alienation and Belonging Experiences
Intense
MA
Dim
High
SR
Low
Routine
MSC
Bizarre
Mental State Clusters (MSC)
Self-Relatedness (SR)
Meta-Awareness (MA)
Problem of Belonging (POB) ●

Fig. 4 Cube Model of Belonging: Note that the *x*- and *y*-axes of Figs. 1 and 3 are here collapsed into one dimension, *MSC*. This is done in order to suggest how the model can accommodate diverse phenomena and how it can handle more than three dimensions

to "hyper-reflexivity," an exaggerated form of "self-awareness." Although I would not characterize this phenomenon in just these terms, the idea of hyper-reflexivity suggests that perhaps experienced bizarreness of lower-level experiences can be made more salient by especially intense degrees of mental awareness. For this example of psychosis then, perhaps it is MA that drives the experience of alienation. On the other hand, for pain asymbolia, perhaps it is the MSC that drives the experience, as is the case with popup perceptual illusions. In other words, the model allows for the empirical possibility of POBs that are driven by either top-down or bottom-up processes.

An additional virtue of this model is that it accounts for both alienation and belonging experiences (Lane et al. 2013). As for the latter, in the case of the rubber hand illusion, subjects become aware of the bizarre clustering of mental states—tactile sensations are experienced as being where they could not be. It is in becoming intensely aware of that "touch referral," which is likely a combination of top-down and bottom-up processes, that subjects begin to experience a rubber hand as belonging to self.

For all types of POB, whether alienation or belonging, MSC, MA, and SR are required. As depicted in the cube model, when MSC is bizarre, MA is intense, and SR is high, a POB will result. Within this conceptual space, the POB is identified by the dot in the rear, upper right-hand corner.

4 Alien Actions and the Explanatory Model of Belonging

Among the several symptoms of patients suffering from schizophrenia are passivity experiences or delusions of alien control. The essence of this symptom is the experience of one's will as having been replaced by some other force or agency (Blakemore et al. 2000b, p. 1132): patients often describe their thoughts, speech, or actions as being controlled by external agents, such as spirits or machines. Such experiences are not necessarily specific for or diagnostic of schizophrenia, but most published reports of these experiences concern people who are suffering from schizophrenia (Nordgaard et al. 2008).

There is no question but that it is the patient's brain or body that *hosts* the thoughts, speech, or actions, but the experience of belonging is disturbed. As regards action, the principal focus of this section, many instances are decidedly unremarkable, concerning such trivial things as combing one's hair or typing on a computer keyboard. Mellor (1970) provides some representative examples: "My fingers pick up the pen, but I don't control them. What they do is nothing to do with me." Frith et al. (2000, p. 18) provide others that include some commonplace sources of alien agency: "My grandfather hypnotized my brain and now he moves my foot up and down." "They inserted a computer in my brain. It makes me turn to the left or right."

It should be noted that many authors who describe these symptoms distinguish between agency and ownership. Hirjack and Fuchs (2010, p. 100), for example, describe a patient who experienced bodily movements as "being made, controlled, and steered by outside forces." Nevertheless, according to Hirjack and Fuchs, the experience of these bodily movements as "belonging to himself was still preserved." In other words, agency is aberrant, but some form of ownership seems intact. This distinction, though commonly invoked in discussions of the experience of schizophrenic passivity experiences, has recently been challenged (Bayne 2010, pp. 156–162; Lane 2012, pp. 279–280; Martin and Pacherie 2013). This debate, however, need not detain us here. It is sufficient that there is a consensus that agency for, or authorship of, the movements is experienced as not belonging to self.

One important respect in which ordinary action differs from alien action is that the experience of action-related sensations is less vivid in the former than in the latter. Frith (2005, p. 752), commenting on this difference, has observed that "the normal mark of the self in action is that we have very little experience of it." But most philosophers and scientists seem to presuppose ordinary action is accompanied by a distinctive sensory experience of some sort. Bayne (2011, p. 356), citing the example of waiting tables, claims that "as you pour the water, you experience yourself as an agent. You experience yourself as someone who is doing something." And, Kühn et al. (2013, p. 1936) claim that we frequently do "make instrumental actions where we have a definite background feeling or buzz of being in control."

Although I incline to the position that ordinary actions are silent,[9] for purposes of the ideas developed here, following Frith, it is sufficient to allow that the conscious

[9] I will not develop my reasons in detail, but I submit that part of the confusion in this vicinity is due to a failure to distinguish clearly and consistently between "qualitative" psychological states, like perceptions and sensations, and "non-qualitative" states, like beliefs or desires (cf. Rosenthal

experience of ordinary and alien action differs. In this spirit, we might then say that ordinarily the phenomenology of self in action is "thin" (Metzinger 2003; Tsakiris et al. 2007, p. 645), while the phenomenology of alien actions is "thick." "Thin" here might, at least in part, be understood as *just knowing* when there is *negligible dwell time:* "when I push a light switch I just 'know' that the light came on because of my action—that I am the author of the action and its consequent effect. There is negligible dwell time…" (Obhi and Hall 2011). Of course when we pause to reflect and render judgment that is a different matter, but here the concern is with what occurs in the nonce, while the action unfolds. And that does not seem to require thick sensory activity. But how is it that some actions are accompanied by a thick phenomenology and why are these felt to be alien?

If the model limned previously is taken as a guide, we can begin to answer these questions by observing that expectations about how mental states should cluster are confounded. The tacit expectation that action phenomenology be thin is confounded. But in virtue of what mechanism is it that what is normally thin becomes thick?

Discussions of ordinary instances of action often begin with the observation that although images on the retina change as a person's eyes move, the visual world is nevertheless experienced as stable (e.g., Churchland 2002, p. 85). Whether eye movement is due to saccades or tracking, the retina registers significant, continuous changes in light patterns. Nevertheless, because our brains interpret these changes as due to the motion of our eyes, rather than motion in the world, stable objects in the world are experienced as remaining stable. On the other hand, when objects in the world do move, our brains can distinguish that from the movement of our eyes.[10] So it is that in all normal circumstances, we can distinguish self-generated movement from movement in the external world—detecting motion when it is there and not confusing it with motion derived from the activity of our eyes. How is it that the brain seems so effortlessly and reliably to make this distinction?

On one view, agents are equipped with "forward action models," so named because they can estimate desired results prior to the realization of actions (e.g., Blakemore et al. 1999; Blakemore et al. 2000a). Succinctly, simultaneous to the sending of a motor command an efference copy (a "corollary discharge") of the command is sent, a copy that makes possible prediction of the movement's sensory consequences. Predicted and actual effects (re-afferences) are then compared. When these two match, the comparator emits a signal indicating that movement is self-generated. In this way, we are able to distinguish between the motion of our eyes and motion in the world.

2005, pp. 218–219, 303–305). Bayne et al., when referring to ordinary action, seem to be writing about non-qualitative states. My position is that in ordinary action the qualitative states though are silent. That is, ordinarily, when we act, our actions are not accompanied by distinctive perceptions or sensations. As to whether ordinary actions are accompanied by distinctive non-qualitative states, that is a separate issue.

[10] The visual experience that objects in the world are moving, when they are actually stationary, can also be caused by gently pressing the eyeball or by paralyzing the eye with curare (Stephens and Graham 2000, p. 136).

This model, as well as others that similarly emphasize congruence between predicted and actual outcomes (Pacherie 2011, p. 448), have been applied to action in general. One important aspect of these models is that they predict when outcomes match they result in "sensory attenuation." That is when predicted sensory consequences occur they are suppressed or attenuated. In other words, in the ordinary course of events, when persons regard themselves as the agent of an action, sensory consequences—proprioceptive and somatosensory sensations—of the action are not vivid. On the contrary, sensory consequences of actions not attributed to self are more vividly felt. "Conscious perception reflects only the error generated by this comparison," the degree to which predicted and actual outcomes fail to match (Voss et al. 2010). Baldly, when expectations are confounded, when proprioceptive and somatosensory sensations are felt, those actions are alien.

Sensory attenuation presupposes another dimension though. In order for this to occur, the behaviors performed or the behaviors observed must be felt to be actions attributable to an agent who harbors intentions. Intended, voluntary actions, unlike unintended, involuntary actions, evince what has come to be known as "intentional binding" (Haggard et al. 2002). "Binding" is of a temporal sort, such that the interval between actions and their consequences is experienced as shorter than is actually the case. In the standard experimental paradigm, the subjective experience of elapsed time between the act of pressing down on a key and the effect of hearing a tone is measured. Act and consequence seem to "attract" one another when the act is felt to be agentive.[11] This time compression does not occur when acts are not felt to be agentive.

Importantly, intentional binding is not specific to self (Wohlschläger et al. 2003). For example, when observing human hands, we temporally bind their movements to the consequences of their movements. Intentional or temporal binding, however, does not occur when what we observe is a rubber hand pulled by a mechanical device. It seems to be the case that intentional binding occurs for both self and other, just so long as the other we observe is a potential agent (cf. Moore et al. 2013). What matters then for intentional binding are the following: (a) a basis for inferring the existence of an intention (e.g., to press a button); (b) the observation of an action consistent with that intention (e.g., actual pressing); and (c) an expected, temporally contiguous sensory effect (e.g., hearing a tone).

Intentional binding has been shown to vary inter-personally. For example, those who suffer from autistic spectrum disorder evince reduced binding, relative to healthy participants (Sperduti et al. 2013). Just the opposite occurs in those who suffer from schizophrenia or in those who have taken ketamine, which when taken at subanesthetic levels causes a state that resembles schizophrenia in certain respects (Moore and Obhi 2012, pp. 554, 555). In effect, those who have passivity experiences belong to a subset of persons who experience "hyper-binding"—actions and their consequences are experienced as even more tightly bound in time than are the actions and consequences of healthy persons who experience intentional binding.

[11] Binding comprises both a predictive and a retrospective mechanism; schizophrenia is associated with an impairment to the former (Voss et al. 2010).

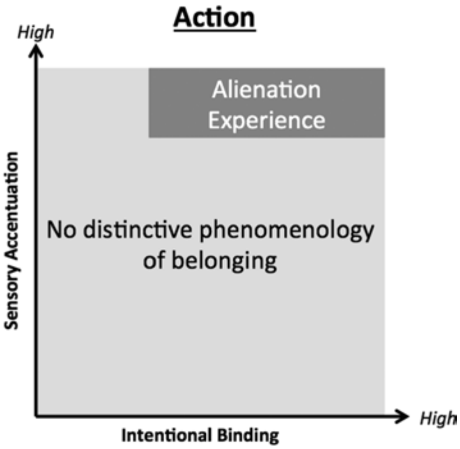

Fig. 5 Principle of confounded expectations: the case of alien actions

What appears to be the case with passivity experiences then is the reverse of what occurs in pain asymbolia. In pain asymbolia, expectations are confounded because something is missing. In passivity experiences, something is added.

For passivity experiences, the feeling that our actions are under the control of some "external force" results when we feel what we should not feel (Frith 2005, p. 753). In ordinary actions sensory attenuation occurs, but here attenuation fails; instead, proprioceptive and somatosensory sensations are heightened. When a person reports that his grandfather is moving his foot up and down, or that a computer inserted into his brain is causing him to turn right or left, because intentional binding applies to both self and others, a sense that observed behaviors are agentive is retained. What differs is the added awareness of the sensory consequences of those actions.

Figure 5 depicts this relationship. Whereas in the normal case sensory attenuation should occur, here sensory *accentuation* occurs. This is indicated by the *y*-axis. The *x*-axis indicates intentional binding, a robust correlate of agency. And since those most likely to have passivity experiences also evince hyper-binding, we can expect that their temporal linking of behavior to its consequences is even stronger than it is for healthy persons. Here, because an atypical mental cluster is realized—sensory accentuation and hyper-binding—a POB in the form of an alienation experience occurs.

Since the attribution of agency is not self-specific, here too we require a *z*-dimension, just as with Fig. 3. Indeed it is the case that there is some evidence that anterior regions of the CMS are involved in the personal experience of agency, especially the anterior cingulate and the medial prefrontal cortex (Northoff and Bermpohl 2004, pp. 102, 103). Accordingly, an adequate explanation of alien actions would seem to require both personal and sub-personal levels: sensory accentuation, hyper-binding, SRP in the anterior CMS, and (see Fig. 4) MA of this bizarre clustering of mental states.

It is not clear whether there are any action-related phenomena that would correspond to the "familiar" or "other" processing indicated in Fig. 3, though the model implies this empirical possibility.[12] But the model, along with some recent investigations and descriptions of other schizophrenic symptoms, suggest that perhaps all passivity experiences can be treated within the framework adumbrated here. Whereas most prior research has focused on agency for action (cf. Frith 2005, p. 753), recent experimental investigations have begun attending to the sense of *agency for thought* (Swiney and Sousa 2013).

There is some evidence to suggest that thought insertions too are accompanied by vivid sensory experiences. Cahill and Frith (1996, p. 278), for example, report that "one of our patients reported *physically feeling* the alien thoughts as they entered his head and claimed that he could *pin-point the point of entry*!"[13] Indeed, Mullins and Spence (2003, p. 295) record that thought insertions "in some (if not all) patients…can incorporate abnormalities of perception." These clinical reports and the explanatory framework adopted here suggest that perhaps passivity experiences in general are characterized by vivid sensory experiences. Such experiences serve as a phenomenological marker of alien agency.

The scope of this hypothesis could be further tested by assessment of the callosal type of anarchic (or, wayward) hand syndrome, which typically results from damage to the anterior corpus callosum (Uddin 2011, p. 97; Verleger et al. 2011). As Marcel (2003, p. 81) has written, for those afflicted with anarchic hand, "there is a positive *otherness* to the anarchic actions." Although the anarchic hand, usually the left, does seem to execute goal-directed movements, patients do not experience these actions as regulated by self. This "otherness" can even be extreme to the point of allowing for inter-manual conflict: for example, a patient puts clothes on with the right hand, while pulling them off with the left (Aboitiz et al. 2003, p. 253; cf., Barbeau et al. 2004).

Anarchic hand does differ from passivity experiences in at least one respect: the alien actions of schizophrenics tend to be attributed to a distinct agent, whereas the alien actions performed by those with wayward hands are not. Although anarchic hands evince contrariness, seemingly a mind of their own, still these patients are not inclined to ascribe those actions to a specific agent. Since the patients lack access to the intentions which motivate actions of the anarchic hands (e.g., Miller et al. 2010), perhaps that is sufficient to explain the experience of "otherness." But lacking access to intention does not entail that patients will experience otherness, and in fact when the central thalamic nucleus is electrically stimulated causing movements that patients describe as goal-directed and voluntary, they simply say they do not know

[12] Perhaps the contents of auditory hallucinations, however, are explainable in such terms (Lane and Northoff 2012). Auditory hallucinations can be the sound of one's own voice, as in "thought broadcast" (Pawar and Spence 2003), or they could be the voices of someone familiar or someone unknown (David 2004). The model proposed here suggests that the difference among these three types might be due to self-related processing. For a novel hypothesis of auditory verbal hallucinations that touches upon some related issues, see, Northoff and Qin (2011) and Northoff (2013a, p. 349).

[13] Italics not contained in original.

why they made those movements (Marcel 2003, pp. 72–73).[14] They do not report experiencing “otherness.”

The framework developed here suggests that to the extent that those who experience anarchic hand describe those actions as alien, those patients should also report sensory *accentuation*. Since they do not suffer from schizophrenia, and thereby, are unlikely to experience hyper-binding, the degree to which their actions are felt to be alien are unlikely to be as vivid as the actions of those who suffer from passivity experiences. But if damage to the anterior corpus callosum does not interfere with SRP, their anarchic actions should correlate with sensory accentuation. If a governing principle of the phenomenology of actions is to be found, it is this: actions are silent unless they are alien.

5 Conclusion

The model depicted in Fig. 4 is still but a toy model. Among other things, in order to make it more comprehensive I have reduced MSC to just one dimension, and prior to that I treated the clustering of mental states as reducible to two dimensions. Even when discussing action, I considered no more than two dimensions—intentional binding and sensory accentuation. But this simplification is nothing more than a heuristic; the confounding of expectations is not likely to be reducible to so few dimensions. A completed model of belonging, or mental glue, will include more than just the dimensions limned here. If the ideas developed previously approximate the truth, they are just a point of departure, not an endpoint.

In the case of action, further attempts to understand alienation will include testing the hypotheses proffered previously. Is it the case that anarchic actions are experienced as anarchic because of an MA of sensory accentuation and intentional binding that are accompanied by SRP? Is it also the case that thought insertion is like other forms of passivity experience, in which it is accompanied by heightened sensory activity? The model and strands of data pieced together from various case studies or experimental investigations suggest as much. But obviously, the evidence adduced here on behalf of these hypotheses falls far short of confirmation.

The search for further confirmation is worth pursuing though, not only because it might enhance our understanding of these pathologies. Beyond that, it might also assist with the development of this toy model into a more mature version of greater scope that can guide research into previously unanticipated domains, which is after all one of the marks of a good theory. There is always risk in such an approach, not least of which is the concern that we might be committing what can be dubbed the “natural kinds fallacy.” Perhaps, the phenomena described previously are only

[14] Marcel does observe that it is possible that stimulation of the thalamus causes both the action and an intention, but patients never reported an “urge to do the action.” This leaves only the possibility that an unconscious intention was triggered by the stimulus, but there is no evidence to support this claim.

superficially similar, no more alike than are gold and iron pyrite. Similarity of subjective report does not entail that the phenomena can all be treated from within a unified framework. But the claim advanced here is that we already have sufficient evidence of a suggestive sort, sufficient to warrant further investigation. And a virtue of the model presented here is that it does generate testable hypotheses, which will stand or fall, depending upon the results of clinical or experimental investigations.

One reason for cautious optimism that this model or some more mature version of it will stand the test of time, is that it is consistent with many developing theories of conscious experience that advocate a deflationary view of its functional role, a view that is compelling because much evidence has accumulated to show that consciousness is slow, capable of just a limited capacity, and unnecessary for many perceptual and cognitive activities (Churchland 2002, pp. 127–200; Edelman et al. 2011; Koch 2004, pp. 205–229; Rosenthal 2008). Rather than becoming involved in the quotidian, consciousness seems much better suited to dealing with novelty or error, and indeed "the unexpected has an especially privileged change of 'gaining access' to consciousness" (Gray 2004, p. 76). As regards the issue at hand, "the unexpected" is constituted by the way in which self-related, atypical clustering of mental states confounds expectations.

How are these seeming limits to consciousness related to experiences of self? They suggest that many who explore this terrain are in error from the get-go. Craig (2009, p. 65), for example, writes: "I regard awareness as knowing that one exists (*the feeling that 'I am'*); an organism must be able to experience its own existence as a sentient being before it can experience the existence and salience of anything else in the environment."[15] The considerations adduced here and elsewhere (Lane 2012) suggest that there is no "feeling that 'I am'," at least that is if this is literally taken to be a condition for experiencing "the existence and salience of anything else in the environment." To repeat just one among the reasons cited above: consciousness is slow. If we had to wait for a "feeling that I am," that wait would be an obstacle to experiencing the existence and salience of anything else.

We do reflect on self. Some of us do this more; some, less. But this is when we can extract ourselves from the hustle and bustle of daily life, and it is not likely well construed as an *experience* of self, for these reflections do not seem to involve or require any obvious sensory experiences. A contention of this essay is that our most direct conscious access to self in the nonce is when we become aware of its absence. Ordinarily SRP takes place on a sub-personal level, out-of-sight, so to speak. It is only when we become aware of self-related mental states clustering in bizarre ways that we become, indirectly, aware of self. And if we wish to learn about self, we would be well advised to attend closely to these absences.

Acknowledgments I express my heartfelt gratitude to Ellie Hua Wang and to Tzu-Wei Hung for their constructive comments on previous versions of this manuscript. For much useful discussion, I am also grateful to the many other participants in Academia Sinica's *International Conference on Language and Action,* Taipei, Taiwan (September 17–18, 2013). Funding for this research was, in part, provided by National Science Council of Taiwan research grants, 100-2410-H-038-009-MY3 and 102-2420-H-038-001-MY3.

[15] Italics not contained in the original.

References

Aboitiz, F., Carrasco, X., Schröter, C., Zaidel, D., Zaidel, E., & Lavados, M. (2003). The alien hand syndrome: Classification of forms reported and discussion of a new condition. *Neurological Sciences, 24*(4), 252–257.

Auvray, M., Myin, E., & Spence, C. (2010). The sensory-discriminative and affective-motivational aspects of pain. *Neuroscience & Biobehavioral Reviews, 34*(2), 214–223.

Baars, B. J. (2007). Attention and consciousness. In B. J. Baars & N. M. Gage (Eds.), *Cognition, brain, and consciousness, second edition: Introduction to cognitive neuroscience* (pp. 225–254). New York: Academic Press.

Baars, B. J., Ramsøy, T. Z., & Laureys, S. (2003). Brain, conscious experience and the observing self. *Trends in Neurosciences, 26*(12), 671–675.

Barbeau, E., Joubert, S., & Poncet, M. (2004). A single case-study of diagnostic dyspraxia. *Brain and Cognition, 54*(3), 215–217.

Bayne, T. (2010). *The unity of consciousness*. New York: Oxford University Press.

Bayne, T. (2011). The sense of agency. In F. Macpherson (Ed.), *The senses: Classic and contemporary philosophical perspectives* (pp. 355–374). New York: Oxford University Press.

Berthier, M., Starkstein, S., & Leiguarda, R. (1988). Asymbolia for pain: A sensory-limbic disconnection syndrome. *Annals of Neurology, 24,* 41–49.

Blakemore, S.-J., Frith, C. D., & Wolpert, D. M. (1999). Spatio-temporal prediction modulates the perception of self-produced stimuli. *Journal of Cognitive Neuroscience, 11*(5), 551–559.

Blakemore, S.-J., Frith, C. D., & Wolpert, D. M. (2000a). Why can't you tickle yourself? *NeuroReport, 11,* R11–R16.

Blakemore, S.-J., Smith, J., Steel, R., Johnstone, E., & Frith, C. (2000b). The perception of self-produced sensory stimuli in patients with auditory hallucinations and passivity experiences: Evidence for a breakdown of self-monitoring. *Psychological Medicine, 30*(5), 1131–1139.

Bor, D. (2012). *The ravenous brain: How the new science of consciousness explains our insatiable search for meaning*. New York: Basic Books.

Bortolotti, L. (2010). *Delusions and other irrational beliefs*. New York: Oxford University Press.

Cahill, C., & Frith, C. (1996). False perceptions or false beliefs? Hallucinations and delusions in schizophrenia. In P. W. Halligan & J. C. Marshall (Eds.), *Method in madness: Case studies in cognitive neuropsychiatry* (pp. 267–291). New York: Psychology Press.

Carruthers, P. (2000). *Phenomenal consciousness: A naturalistic theory*. New York: Cambridge University Press.

Churchland, P. S. (2002). *Brain-wise: Studies in neurophilosophy*. Cambridge: The MIT Press.

Clark, A., Anderson, M. L., Block, N., Bowman, H., Bridgeman, B., Buckingham, G., et al. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.

Craig, A. D. (2009). How do you feel—now? The anterior insula and human awareness. *Nature Reviews: Neuroscience, 10*(1), 59–70.

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nature Neuroscience, 6*(2), 119–126.

David, A. S. (2004). The cognitive neuropsychiatry of auditory verbal hallucinations: An overview. *Cognitive Neuropsychiatry, 9*(1/2), 107–123.

Davies, M. (2000). Interaction without reduction: The relationship between personal and subpersonal levels of description. *Mind & Society, 1*(2), 87–105.

de Pinedo-Garcia, M., & Noble, J. (2008). Beyond persons: Extending the personal/subpersonal distinction to non-rational animals and artificial agents. *Biology and Philosophy, 23*(1), 87–100.

Edelman, G. M., Gally, J. A., & Baars, B. J. (2011). Biology of consciousness. *Frontiers in Psychology, 2,* 1–7.

Ellis, H. D., & Lewis, M. B. (2001). Capgras delusion: A window on face recognition. *Trends in Cognitive Sciences, 5*(4), 149–156.

Feinberg, T. E. (2009). *From axons to identity: Neurological explorations of the nature of the self.* New York: W. W. Norton.

Fitzgibbon, B. M., Giummarra, M. J., Georgiou-Karistianis, N., Enticott, P. G., & Bradshaw, J. L. (2010). Shared pain: From empathy to synaesthesia. *Neuroscience and Biobehavioral Reviews, 34*(4), 500–512.

Frith, C. (2005). The self in action: Lessons from delusions of control. *Consciousness and Cognition, 14*(4), 752–770.

Frith, C. D., Blakemore, S.-J., & Wolpert, D. M. (2000). Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews, 31*(2/3), 357–363.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Science, 13*(7), 293–301.

Gallagher, S. (2012). First-person perspective and immunity to error through misidentification. In S. Miguens & G. Preyer (Eds.), *Consciousness and subjectivity* (pp. 187–214). Heusenstamm: Ontos.

Grahek, N. (2007). *Feeling pain and being in pain* (2nd ed.). Cambridge: The MIT Press.

Gray, J. (2004). *Consciousness: Creeping up on the hard problem*. New York: Oxford University Press.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience, 5*(4), 382–385.

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: The Free Press.

Hirjack, D., & Fuchs, T. (2010). Delusions of technical alien control: A phenomenological description of three cases. *Psychopathology, 43*(2), 96–103.

Hohwy, J. (2013). *The predictive mind*. New York: Oxford University Press.

Jones, E. G. (2007). *The thalamus* (2nd ed.). New York: Cambridge University Press.

Klein, S. B. (2013a). *The two selves: Their metaphysical commitments and functional independence*. New York: Oxford University Press.

Klein, S. B. (2013b). Making the case that episodic recollection is attributable to operations occurring at retrieval rather than to content stored in a dedicated subsystem of long-term memory. *Frontiers in Behavioral Neuroscience, 7*(3), 1–14. doi:10.3389/fnbeh.2013.00003.

Klein, S. B., & Nichols, S. (2012). Memory and the sense of personal identity. *Mind, 121*(483), 677–702.

Koch, C. (2004). *The quest for consciousness: A neurobiological approach*. Englewood: Roberts and Company.

Kühn, S., Brass, M., & Haggard, P. (2013). Feeling in control: Neural correlates of experience of agency. *Cortex, 49*(7), 1935–1942.

Lane, T. (2012). Toward an explanatory framework for mental ownership. *Phenomenology and the Cognitive Sciences, 11*(2), 251–286.

Lane, T., & Liang, C. (2008). Higher-order thought and the problem of radical confabulation. *Southern Journal of Philosophy, 46*(1), 69–98.

Lane, T., & Liang, C. (2010). Mental ownership and higher-order thought. *Analysis, 70*(3), 496–501.

Lane, T., & Liang, C. (2011). Self-consciousness and immunity. *Journal of Philosophy, 108*(2), 78–99.

Lane, T., & Northoff, G. (2012, July). Mineness, Minimal Self, and Self-Related Processing. Paper presented at the 16th Annual Meeting of the Association for the Scientific Study of Consciousness. Brighton, UK.

Lane, T., Yeh, S., & Chang, A. (2013, July). *Switching Attention to the Rubber Hand*. Paper presented at the 17th Annual Meeting of the Association for the Scientific Study of Consciousness, San Diego, CA.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences, 15*(8), 365–373.

Marcel, A. (2003). The sense of agency: Awareness and ownership of action. In J. Roessler & N. Eilan (Eds.), *Agency and self-awareness: Issues in philosophy and psychology* (pp. 48–93). Oxford: Clarendon.

Martin, J.-R., & Pacherie, E. (2013). Out of nowhere: Thought insertion, ownership and context-integration. *Consciousness and Cognition, 22*(1), 111–122.

McDowell, J. (1994). *Mind and world*. Cambridge: Harvard University Press.

Mellor, C. S. (1970). First-rank symptoms of schizophrenia. *British Journal of Psychiatry, 117*(536), 15–23.

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Cambridge: The MIT Press.

Miller, M. B., Sinnott-Armstrong, W., Young, L., King, D., Paggi, A., Fabri, M., et al. (2010). Abnormal moral reasoning in complete and partial callosotomy patients. *Neuropsychologia, 48*(7), 2215–2220.

Moore, J. W., & Obhi, S. S. (2012). Intentional binding and the sense of agency: A review. *Consciousness and Cognition, 21*(1), 546–561.

Moore, J. W., Teufel, C., Subramaniam, N., Davis, G., & Fletcher, P. C. (2013). Attribution of intentional causation influences the perception of observed movements: Behavioral evidence and neural correlates. *Frontiers in Psychology, 4,* 23. doi:10.3389/fpsyg.2013.00023.

Mullins, S., & Spence, S. A. (2003). Re-examining thought insertion: Semi-structured literature review and conceptual analysis. *British Journal of Psychiatry, 182,* 293–298.

Nordgaard, J., Arnfred, S. M., Handest, P., & Parnas, J. (2008). The diagnostic status of first rank symptoms. *Schizophrenia Bulletin, 34*(1), 137–154.

Northoff, G. (2013a). *Unlocking the brain: Vol. 2. Consciousness*. New York: Oxford University Press.

Northoff, G. (2013b). Unlocking the brain: Vol. 1. Coding. New York: Oxford University Press.

Northoff, G., & Bermpohl, F. (2004). Cortical midline structures and the self. *Trends in Cognitive Sciences, 8*(3), 102–107.

Northoff, G., & Panksepp, J. (2008). The trans-species concept of self and the subcortical-cortical midline system. *Trends in Cognitive Science, 12*(7), 259–264.

Northoff, G., & Qin, P. (2011). How can the brain's resting state activity generate hallucinations? A "resting state hypothesis" of auditory verbal hallucinations. *Schizophrenia Research, 127,* 202–214.

Northoff, G., Lane, T., & Yen, N. (2014). *Pre-stimulus neural activity predicts self-relatedness judgments in healthy subjects*: *A multi-modal study*. Unpublished manuscript.

Pacherie, E. (2011). Self-agency. In S. Gallagher (Ed.), *The Oxford handbook of the self* (pp. 442–464). Oxford: Oxford University Press.

Pawar, A. V., & Spence, S. A. (2003). Defining thought broadcast: Semi-structured literature review. *British Journal of Psychiatry, 183,* 287–291.

Prinz, J. (2012a). Waiting for the self. In J.-L. Liu & J. Perry (Eds.), *Consciousness and the Self: New essays* (pp. 123–149). New York: Cambridge University Press.

Prinz, J. (2012b). *The conscious brain: How attention engenders experience*. New York: Oxford University Press.

Obhi, S. S., & Hall, P. (2011). Sense of agency and intentional binding in joint action. *Experimental Brain Research, 211*(3/4), 655–662.

Ochsner, K. N., Zaki, J., Hanelin, J., Ludlow, D. H., Knierim, K., Ramachanran, T., et al. (2008). Your pain or mine? Common and distinct neural systems supporting the perception of pain in self and other. *Social Cognitive and Affective Neuroscience, 3*(2), 144–160.

Qin, P., & Northoff, G. (2011). How is our self related to midline regions and the default-mode network? *NeuroImage, 57*(3), 1221–1233.

Raichle, M. E. (2010). Two views of brain function. *Trends in Cognitive Science, 14*(4), 180–190.

Rosenthal, D. M. (2002). Explaining consciousness. In D. J. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 406–421). New York: Oxford University Press.

Rosenthal, D. M. (2005). *Consciousness and mind*. New York: Oxford University Press.

Rosenthal, D. M. (2008). Consciousness and its function. *Neuropsychologia, 46*(3), 829–840.

Sass, L. A., & Parnas, J. (2003). Schizophrenia, consciousness, and the self. *Schizophrenia Bulletin, 29*(3), 427–444.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Science, 15*(7), 319–326.

Seth, A. K., Suzuki, K., & Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology, 2,* 395.

Shea, N. (2013). Neural mechanisms of decision-making and the personal level. In K. W. M. Fulford, M. Davies, R. G. T. Gipps, G. Graham, J. Z. Sadler & G. Stanghellini (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 1063–1082). New York: Oxford University Press.

Shoemaker, S. (1994). Self-knowledge and "inner Sense": Lecture I: The object perception model. *Philosophy and Phenomenological Research, 54*(2), 249–269.

Sierra, M. (2009). *Depersonalization: A new look at a neglected syndrome*. New York: Cambridge University Press.

Sorensen, R. (2007). The vanishing point: A model of the self as an absence. *Monist, 90*(3), 432–456.

Sperduti, M., Pieron, M., Leboyer, M., & Zalla, T. (2013, July 24). Altered pre-reflective sense of agency in autism spectrum disorders as revealed by reduced intentional binding. *Journal of Autism and Developmental Disorders*, *44*(2), 343–352. doi:10.1007/s10803–013-1891-y. (http://link.springer.com/article/10.1007/s10803–013-1891-y.)

Stephens, G. L., & Graham, G. (2000). *When self-consciousness breaks: Alien voices and inserted thoughts*. Cambridge: The MIT Press.

Stone, J. L., & Goodrich, J. T. (2006). The craniopagus malformation: Classification and implications for surgical separation. *Brain, 129*(5), 1084–1095.

Swiney, L., & Sousa, P. (2013). When our thoughts are not our own: Investigating agency misattribution using the Mind-to-Mind paradigm. *Consciousness and Cognition, 22*(2), 589–602.

Tsakiris, M., Schütz-Bosbach, S., & Gallagher, S. (2007). On agency and body-ownership: Phenomenological and neurocognitive reflections. *Consciousness and Cognition, 16*(3), 645–660.

Uddin, L. Q. (2011). Brain connectivity and the self: The case of cerebral disconnection. *Consciousness and Cognition, 20*(1), 94–98.

Verleger, R., Binkofski, F., Friedrich, M., Sedlmeier, P., & Kömpf, D. (2011). Anarchic-hand syndrome: ERP reflections of lost control over the right hemisphere. *Brain and Cognition, 77*(1), 138–150.

Voss, M., Moore, J., Hauser, M., Gallinat, J., Heinz, A., & Haggard, P. (2010). Altered awareness of action in schizophrenia: A specific deficit in predicting action consequences. *Brain, 133*(10), 3104–3112.

Wegner, D. (2002). *The illusion of conscious will.* Cambridge: The MIT Press.

Wohlschläger, A., Haggard, P., Gesierich, B., & Prinz, W. (2003). The perceived onset time of self- and other-generated actions. *Psychological Science, 14*(6), 586–591.

Zahn, R., Talazko, J., & Ebert, D. (2008). Loss of the sense of self-ownership for perceptions of objects in a case of right inferior temporal, parieto-occipital and precentral hypometabolism. *Psychopathology, 41*(6), 397–402.

Zihl, J., von Cramon, D., & Mai, N. (1983). Selective disturbance of movement vision after bilateral brain damage. *Brain, 106*(2), 313–340.

Self-Consciousness and Its Linguistic Expression

Rory Madden

Abstract Which linguistic actions are expressions of self-conscious states of mind? I defend a certain answer to this question. Having presented problems for a simple view of the connection between self-conscious states of mind and first person language, and for a slight modification of the view, I go on to distinguish two, more promising, ways of getting a linguistic handle on first person thought. These two positions—which I call the Knowledge View and the Intention View—are not explicitly distinguished in the existing literature on the subject. My aim is to argue that the Intention View is the superior view. One reason for preferring the Intention View is its capacity to furnish a noncircular route to the identification of first person thoughts. This advantage accrues from the way in which objects of intention contrast with objects of propositional knowledge. Another reason for preferring the Intention View is that it diagnoses what is going on in certain persuasive counter-examples to the Knowledge View. In the final section of this chapter, I consider whether the Intention View is subject to some counter-examples of its own. Clarification of the relevant notion of linguistic expression reveals the counter-examples to be merely apparent.

1 Introduction

How is self-consciousness expressed in language? Which utterances are expressions of self-conscious states of mind? I want to defend a certain answer to this question.

Here is a familiar example to illustrate the notion of a self-conscious state of mind.

> As John rounds the aisles of the supermarket he spies a trail of spilled sugar, and comes to believe that the shopper with a torn bag is making a mess. In fact John is unwittingly depositing the trail of sugar from a torn bag in his own supermarket trolley. After some minutes of futile pursuit of the shopper—the trail of sugar growing thicker as he goes round and round the same shelves—it finally dawns on John that he himself is making the mess. He stops the trolley and rearranges the bag.

R. Madden (✉)
Department of Philosophy, University College London, London, UK
e-mail: r.madden@ucl.ac.uk

Tzu-Wei Hung (ed.), *Communicative Action,* DOI 10.1007/978-981-4585-84-2_5,

Since John himself was the shopper with the torn bag all along his belief that the shopper with a torn bag is making mess was, in fact, a belief about himself. So he was in one sense thinking about himself from the outset. When finally he comes to realize that he himself is making a mess he is still thinking about himself, but there has been a shift in the functional role of his psychological state. His belief now disposes him to stop pushing the trolley and to begin fixing his possessions. He feels embarrassed. He resolves to be more careful in future. He has entered a state one is disposed to enter upon tracing a trail of sugar back to roughly the origin of one's own egocentric representations of the world. States of mind with this distinctive, self-conscious, kind of functional role philosophers are apt to call "first personal" states of mind. After his realization, John thinks of himself first personally.

But the term "first person" primarily effects a *grammatical* categorization, of pronouns, possessive determiners, and verb forms. Why should philosophers use a grammatical term to classify states of mind?

Evidently, philosophers' use of the term here is metonymic; the states of mind in question are states of a kind that are expressively associated with language grammatically categorized as first personal. For example, after his realization, and not before, John will be disposed to express his state of mind using the first person pronoun. As an English speaker, he will now be disposed to say such things as "I am making a mess."

But how straightforward is the connection between first personal states of mind and uses of the first person pronoun and cognate expressions? It is sometimes assumed that first personal states of mind can simply be identified with those states of mind that would be expressed using the first person. This simple view, as I shall shortly argue, is mistaken. In order to more accurately identify first personal states of mind by their means of linguistic expression, we need to keep in mind that expressive language use is a rational, intentional, activity. I shall argue that first personal states of mind are states of mind expressible by linguistic actions performed with a certain intention. This position explains why first personal states of mind correlate imperfectly, though still closely, with uses of first personal language.

In the next section, I present problems for the simple view of the connection between first personal states of mind and first personal language, and for a slight modification of the view. I go on to distinguish two, more promising, ways of getting a linguistic handle on first person thought. These two positions—which I shall call the Knowledge View and the Intention View—are not explicitly distinguished in the existing literature on the subject. My aim is to argue that the Intention View is the superior view. One reason for preferring the Intention View is its capacity to furnish a noncircular route to the identification of first person thoughts. This advantage accrues from the way in which objects of intention contrast with objects of propositional knowledge. Another reason for preferring the Intention View is that it diagnoses what is going on in certain persuasive counter-examples to the Knowledge View. In the final section, I consider whether the Intention View is subject to some counterexamples of its own. Clarification of the relevant notion of linguistic expression reveals the counter-examples to be merely apparent.

2 A Simple View

It is helpful first to reflect on the shortcomings of a simple view of the connection between first personal states of mind and first personal language.

(Simple View) One linguistically expresses a first person thought if and only if one uses first personal language.

The Simple View faces counterexamples, to both the necessity and the sufficiency of first personal language.

Against necessity it is enough to point out that first personal states of mind can be expressed by the arch, or pompous, use of one's own name, à la General De Gaulle (certainly in contexts in which it is common knowledge that it is one's name). Equally, one might use more complex descriptive expressions to give voice to first personal states of mind, such as "the subject of these experiences" or "this speaker." First personal devices are not necessary for the linguistic expression of first person thinking.

Is it promising to regard the use of first personal language as at least a *sufficient* condition for the expression of first person thought?

Some nonstandard uses of "I" suggest otherwise. Suppose that a helpful PA has the bright idea of recording a computerized answering machine message on behalf of The Boss, who is presently incommunicado on a delayed flight. The PA, a competent English speaker, chooses to produce the sentence "I am away from my office."[1] The PA thereby comprehendingly uses first personal language. But the PA does not thereby express a self-conscious, first personal, state of mind. If it is right to say that the PA expresses a thought at all, the PA expresses a "third personal" thought, which would in more normal circumstances be expressed by "The Boss is away from his office." The use of first personal language is not a sufficient condition for the expression of first person thought.

A notable feature of the case is that the PA does not use "I" to refer to himself. It might be thought that the simple view can be adapted accordingly, to give a sound criterion for the expression of first person thought:

(Simple View) One linguistically expresses a first person thought if one uses first personal language *and one thereby refers to oneself.*

The counterexample can be adapted accordingly. In a variant case it is not the PA who records the message but The Boss himself. For The Boss suffers an amnesiac-delusional episode on the airplane, during which he forgets entirely who he is and comes to believe that he is the earthbound PA to the airborne Boss. He uses a satellite phone to record a first personal message "I am away from my office"—as he conceives of it, on behalf of someone else.

[1] Romdenh-Romluc (2008) gives roughly this example, in the service of a general skepticism about using language to identify first person thought. The conclusions of the present essay imply that such skepticism is unwarranted.

The Boss is the speaker's intended referent of "I." The Boss is the person whom listeners of the message will take to be the referent of "I." The Boss is the producer of the word "I." So it is extremely plausible that The Boss thereby refers to himself. However, he no more expresses a first person thought than the PA in the original case. Thus, one can use first personal language and thereby refer to oneself, and yet not express a first person thought. The Simple View* is no improvement over The Simple View.

Can we do better? In her essay "The First Person," G. E. M. Anscombe considers, only to reject, the following proposed explanation of the meaning of the English first person pronoun.

> "I" is the word each one uses when he knowingly and intentionally speaks of himself. (Anscombe 1975, p. 22)

As it stands, this proposal is vulnerable to the examples just considered. In De Gaulle-type cases, the speaker knowingly and intentionally speaks of himself but he does not use the word "I." Conversely, in both the PA cases and their variants the speaker uses "I" but does not knowingly and intentionally speak of himself.

These observations, however, strongly suggest an improved approach to the connection between first person states of mind and their linguistic expression: one linguistically expresses a first person thought if and only if one knowingly and intentionally speaks (writes, etc.) of oneself. None of the recent counterexamples threatens this way of getting a linguistic handle on self-conscious states of mind. Just in so far as he knows that "De Gaulle" is his own name and intends to speak of himself, De Gaulle is using his name to express his first person state of mind. Equally, a speaker who expresses first person states of mind by means of "the subject of these experiences" is a speaker who also thereby knowingly and intentionally refers to himself by means of that descriptive device. We can also note that the approach accords nicely with the opening messy shopper case. Before the onset of his self-conscious belief John will in fact refer to himself, with the expression "the shopper with a torn sack"; what is notable is that he does not refer to himself knowingly and intentionally. When self-awareness dawns he switches to the first person pronoun "I" and now he knowingly and intentionally refers to himself.

I think that something along these lines is correct. However Anscombe's wording "knowingly and intentionally" fuses two strands of thought. Separating these strands yields two views of the connection between first person thought and language:

(Knowledge View) One linguistically expresses a first person thought if and only if one uses language in the knowledge that one is thereby referring to oneself.

(Intention View) One linguistically expresses a first person thought if and only if one uses language with the intention thereby to refer to oneself.

These views are not obviously incompatible. For all the examples so far have shown they are both true.[2]

What I want to argue, though, is that the Knowledge View is incorrect. The Intention View is much more promising. In the next part of the chapter, I shall explain an explanatory advantage of the Intention View over the Knowledge View. I shall then turn to more straightforward counter-examples to the Knowledge View, which the Intention View can handle quite comfortably.

3 Objects of Knowledge and Objects of Intention

Why did Anscombe reject the explanation of the first person pronoun “I” as “the word each one uses when he knowingly and intentionally speaks of himself?” I shall briefly summarize her extended argument, which can be seen as posing a trilemma. This argument is prima facie threatening to both the Knowledge View and the Intention View.

First, note that the ascriptions of knowledge and intention on the right-hand sides of the Knowledge View and the Intention View, must, in order for the two theses to have any plausibility whatsoever, be understood as opaque. In particular, the occurrence of the reflexive pronoun “oneself” after the attitude words cannot be understood as a transparent reference to the speaker, substitutable *salva veritate* with any designation of the speaker. For, on a transparent reading of the ascription, it is true that the messy shopper, even before he enters a first personal state of mind, knows that he refers to, and intends to refer to, himself. For he knows that he refers to, and he intends to refer to, the messy shopper—and the messy shopper is he.

On the other hand, if the attitude ascriptions are opaque, then how should their truth conditions be construed? A natural thought is that the occurrence of “oneself” is an indirect report of a first personal reference made by the speaker to whom the attitude is ascribed. The ascription of knowledge is true only if the speaker knows that he refers to the speaker, while conceiving of the speaker under a first personal guise, or mode of presentation. Similarly, the ascription of intention is true only if the speaker intends to refer to the speaker, conceiving of the speaker under a first personal guise or mode of presentation. With the ascriptions read in this way the Knowledge and Intention Views accord with the datum that the messy shopper does not express a first personal state of mind before his realization. The messy shopper knows that he refers to, and intends to refer to, what is in fact himself—but he does not conceive of the object of reference in first personal terms.

[2] Those attracted to the Knowledge View include O’Brien, who holds that “reflexive reference can only be first-person reference if one knows that one is referring to oneself” (2007, p. 9), and with qualifications, Peacocke: “fully self-conscious uses of ‘I’ are those in which the thinker knows that he is referring to himself” (2008, p. 78). Nozick (1980, p. 79) is another advocate. Evans appears to be sympathetic to both the Knowledge and the Intention Views (1982, pp. 258–259). Although Rumfitt’s (1994) discussion of conventional linguistic meaning of “I” is suggestive, I am aware of no explicit defense of the Intention View. The present essay aims to remedy this situation.

In response to this natural thought, however, one might well follow Anscombe in raising a question about the noncircularity, or explanatory power of the views read in this way. For once driven to understand the relevant occurrences of "oneself" in essentially first personal terms, it seems we can no longer give an informative linguistic identification of first person thought in independent terms. We are distinguishing cases of the expression of first person thought from other cases precisely by reference to first person thought on behalf of the subject.[3]

Anscombe assumes that the only avenue of escape from this circularity is to embark upon the quest for some independently specifiable mode of presentation, or conceptual guise, by means of which the speaker singles himself out in the content of the knowledge or intention in question. Anscombe is skeptical about the prospects for such a quest. Neither descriptive nor demonstrative modes of presentation seem adequate. In cases of total descriptive ignorance or error about one's properties—and in sensory deprivation cases in which one is in no position to demonstrate oneself—one can surely still knowingly and intentionally refer to oneself in the way that is expressive of first person states of mind.

Thus, the advocate of the general proposal that first personal thoughts are those linguistically expressible using language by means of which one knowingly and intentionally refers to oneself faces a trilemma. (a) If the occurrence of "oneself" is understood to make a transparent reference to the subject then the proposal is clearly extensionally inadequate. (b) If the occurrence of "oneself" ascribes a first personal way of thinking of the subject then the proposal is extensionally adequate but circular. (c) If the occurrence of "oneself" ascribes a nonfirst personal way of thinking of the subject then, again, the proposal is extensionally inadequate.

How should one respond to this trilemma? While it would take a lot more argument than can be given here, I think Anscombe is certainly right to be skeptical about the possibility of locating some independently specifiable, descriptive or demonstrative mode of presentation of the subject. The weak point in her argument is rather a presupposition of the whole trilemma. In presenting the choice between interpreting the relevant occurrence of "oneself" as a transparent reference to the subject, a first personal reference to the subject, or some other reference to the subject, Anscombe is clearly presupposing that the reflexive pronoun "oneself" within the scope of the attitude verbs has a referential function of some kind.[4] There is a better way of understanding the occurrence of the reflexive pronoun. The right-hand sides of the Knowledge View and the Intention View are ascribing attitudes to a *reflexive act*.

To get a feel for this view, consider the following example. When on a particular occasion Alice washes Alice there is whole range of things that Alice thereby does. She uses water. She uses soap. She washes Alice. She washes someone. The single event in which she does all these things we can call an *action*, which is an unrepeatable event, occurring at a particular time, with a particular agent. The various things she *does*, we can call *acts*. Acts are not unrepeatable particulars like action.

[3] This is the "paradox of self-consciousness" which animates Bermudez's eponymous 1998.

[4] I owe this astute diagnosis to Rumfitt (1994).

Different agents at different times may engage in the very same act. For example, when on some other occasion *Bob* washes Alice, he does one of the very same things that Alice does on this occasion: he performs the act of washing Alice. Acts may be specified by predicates, or open sentences, such as "*x* washes Alice." The predicate specifies the act any *x* engages in when *x* washes Alice.[5]

Now suppose on a particular occasion that Bob washes Bob. There is something Bob does on this occasion that Alice does on an occasion that Alice washes Alice. This act we specify using the *reflexive* predication "*x* washes *x*." The act is the act any *x* engages in when *x* washes *x*. We may call this act the act of self-washing or reflexive washing.

The way out of Anscombe's trilemma, then, is to understand the ascriptions of knowledge and intention as ascribing to the speaker an attitude toward an act of just this sort. It is the act specified by the predicate "*x* refers to *x*": the act of self-referring or reflexively referring. In order to think of such an act, the speaker need to employ no first personal concept: They need only to be able to discern the common feature present in such situations as the situation in which Alice refers to Alice, Bob refers to Bob, Carol refers to Carol (and absent from the situation in which Alice refers to Bob). When the Knowledge and Intention Views ascribe attitudes toward referring to oneself, the "oneself" here is not making a reference, first personal or otherwise. The views ascribe to the speaker attitudes toward the act any *x* does when *x* refers to *x*.

At this point, however, we can begin to discern an advantage of the Intention View over the Knowledge View. The Knowledge View claims that a first person thought is expressed just in case the one knows that one thereby refers to oneself. Let it be agreed that the occurrence of the reflexive pronoun "oneself" is not to be understood as making a reference to the speaker; the condition is that the speaker knows that he thereby performs the act any *x* performs when *x* refers to *x*. An obvious question remains: how should we understand the "one" that heads the sentential complement of the attitude verb?

Here, I think, there is no choice but to regard the "one" as referential. Thus Anscombe's trilemma can again be pressed, now with full force. First, if the "he" is understood transparently then the Knowledge View is false. A complex example is needed to illustrate the point.

Again, John is in fact the messy shopper although he has not yet realized this. John expresses a third personal thought using the words "the messy shopper is reflexively referring." He says this because he lip-reads the words of an evidently messy shopper using the term "the messy shopper." He knows by visual means that the messy shopper performs the act of reflexively referring: He knows that the messy shopper refers to the messy shopper. What he does not know is that he is in fact looking at himself in a mirror. Thus, with the "he" understood to make a transparent reference, it is true that John knows that *he* is reflexively referring. For he knows that the messy shopper is reflexively referring—and the messy is shopper is he. But John does not linguistically express a first person thought.

[5] See Rumfitt (1994) for a more detailed presentation of this way of identifying acts.

It is clear then, that for the Knowledge View to have plausibility, the "he" heading the complement of the verb "knows" cannot be transparent; indeed the view is surely only plausible if the "he" ascribes a first personal thought, which would be expressed by the words "I am referring to myself." The nonreferential reading of the original reflexive does not defuse this point. Thus, the Knowledge View remains accused of helping itself to the ascription of first person thought in identifying cases of the linguistic expression of first person thought.

Why, though, should the Intention View be supposed to have any advantage here? Does not exactly the same problem arise for the Intention View?

The same problem would arise just in case one made the admittedly common "propositionalist" assumption that the objects of intention, like the objects of propositional knowledge, must be propositional in nature. On this common assumption, an intention to ϕ must be understood as an intention *that* one will ϕ. In the present case, the view would have it that the intention to reflexively refer must be understood as the intention that one will reflexively refer. If this propositionalist view of the objects of intention is correct, then the same difficult question arises for the Intention View, how to understand the occurrence of "one" heading the sentential complement of the intention ascription. It is again plausible that the view will be false unless the "one" is understood as making a first personal reference to the speaker. So no less than the Knowledge View, the Intention View must help itself to the ascription of first person thought in order to identify first person thoughts by their conditions of linguistic expression.

Now, one reaction to this would be to question the supposed viciousness of this element of circularity. One might still find the Knowledge and the Intention Views interesting and informative claims even if they fail to provide an impeccably noncircular route to the identification of first person states of mind via their linguistic expression.

But in fact an advocate of the Intention View need not retreat to this less ambitious stance. Instead the advocate of the Intention View may refuse to accept that an intention to ϕ must be understood as an attitude to a proposition in the first place. The alternative view is to regard the objects of intention as simply *acts*. Whereas, the things we know are such things as *that the Battle of Hastings was in 1066,* the things we intend are naturally regarded as belonging to the same category as the things we do. And the things we do are simply acts, such as washing Alice, buying a ticket, or, as the case may be, reflexively referring. On this view there need be no first personal element to the object of intention itself. While perhaps intentions to act are always in some sense "directed" or "sent" to oneself, a reference to oneself need be no part of the *content* of what is intended.

An analogy with memory may be helpful. A report of *episodic* memory, of the form "*S* remembers ϕ-ing" does not on the face of it appear to ascribe an attitude to a proposition. There is such a thing as propositional first-person memory, ascribed by such reports as "*S* remembers that he himself ϕ-ed" but there is no evident reason to assimilate the former to the latter. A more plausible view is that episodic memory relates the subject not to a proposition but to a past experiential *event*.[6] As things

[6] See Martin (2001) for defense of this view of episodic memory.

are, we only stand in this remembering relation to events in our own past. For this reason, episodic memory is an immediate basis for first person statements of the form "I ϕ-ed." But this does not show that the memory itself has a propositional object, any more than the fact that awareness of a pain is an immediate basis for the statement "I am in pain" shows that the object of awareness is a first person proposition, and not simply a pain.

Although I am myself attracted to an act-directed view of the objects of intention, this is not the place for its full defense. So I cannot claim here to have conclusively shown that the Intention View has the upper hand over the Knowledge View in responding to Anscombe's worries about circularity.[7] Fortunately, there is a far more straightforward way to demonstrate the advantage of the Intention View of the conditions of linguistic expression of first person states of mind. As I shall argue in the next section, there are counterexamples to the Knowledge View. The Intention View handles these cases quite elegantly.

4 Problems for the Knowledge View

There is a tradition in the philosophy of action, inspired by Anscombe's *Intention*, which takes the view that intentional actions are constitutively those events in which the agent knows, by a certain means, what she is doing.[8] A theorist in this tradition will have little inclination to prise apart the "knowingly" and "intentionally" strands in the way I have here.

There is, I think, a general difficulty with the claim that intentionally ϕ-ing requires knowingly ϕ-ing. It is of course perfectly normal in intentionally doing something to know what one is doing. Indeed, it would be practically near impossible to carry through an extended course of intentional action in the absence of successful self-monitoring. But the strong constitutive claim struggles with the point that the conditions for knowing that one is ϕ-ing seem to be *stronger* than the conditions for intentionally ϕ-ing. Here is an example to illustrate the point.

> Scott's toes are totally numbed by the cold and hidden from sight in his hiking boots. Scott believes that he will wiggle his toes whenever he tries. He is wrong: his toes are so paralyzed by cold that the chance that he will wiggle his toes if he tries is extremely low; the chance of a motor signal making it all the way to his toes is less than one percent. On a certain occasion Scott tries to wiggle his toes and, against the odds, he succeeds. So he intentionally wiggles his toes.

[7] A full defense would need to engage with the position in theoretical linguistics, supported by high-level syntactic considerations, that the underlying syntactic structure of the infinitival clause of an intention ascription contains a covert unpronounced subject-term "PRO," and thereby ascribes a complete first personal propositional thought about the intender. A defender of the present response to Anscombe's circularity worries should question whether the underlying syntactic form of a mental state ascription reflects in any straightforward way the metaphysics of the objects of the state.

[8] Anscombe (1957); Velleman (1989); Setiya (2003).

Does he know that he is wiggling his toes? He believes that he is wiggling his toes, and he is right. But in very similar close possible situations he would have been mistaken in his belief that he is wiggling his toes. So it is plausible that he does not know that he is wiggling his toes. He has a merely luckily true belief. Similar examples can be used to elucidate the superiority of the Intention View to the Knowledge View.

> Gertrude is completely deaf, and her vocal apparatus—chest, throat, mouth, and lips—is completely numb. Although she does not know it, she is also paralysed in such a way that attempts to vocalize will in fact succeed only very rarely. Gertrude thinks that she may have overdone the Novocain. As an English speaker she tries to give voice to this thought, and against all the odds, succeeds, uttering the sentence "I may have overdone the Novocain".

It is plausible that Gertrude gives linguistic expression to her self-conscious state of mind. According to the Knowledge View, then, she must be using language in the knowledge that she is thereby self-referring. However, while her belief that she is self-referring by means of "I" is true, it is not plausible that she knows that she is self-referring by means of "I." Given the improbability of motoric success, and the absence of any alternative auditory or proprioceptive means of self-monitoring, Gertrude's belief is extremely unsafe. She could very easily in close possible situations have been mistaken in believing that she is self-referring by means of "I." So she does not know that she is self-referring. She has a merely luckily true belief to that effect.

The case is quite unlike the situation of the ignorant messy shopper who uses "the messy shopper" without knowing that he is thereby self-referring. Unlike the ignorant messy shopper Gertrude linguistically expresses a first person state of mind. The Intention View respects this difference. Gertrude, unlike the messy shopper, speaks with the *intention* to self-refer.

Could the defender of the Knowledge View individuate methods in such a way that basic motoric success is required for the method distinctive of Anscombian "practical knowledge" genuinely to be operative? On the basis of this externalist individuation of methods, it might be argued that the close possible cases of error, while subjectively similar to the actual case of successful wiggling, cannot be cases in which the same method is used to form the belief that one is ϕ-ing.[9] The safety of Gertrude's true belief that she is self-referring, based on one method, is not undermined by the close possibility of making a mistake on the basis of some distinct method.

However this may be, the shortcomings of the Knowledge View can be illustrated with examples in which there is no such breakdown in basic motor control.

> Edmund has a shaky grasp of English vocabulary. He defers to the English linguistic community in his use of English words, intending them to have

[9] See Nozick's "Jesse James" and "Sick Grandmother" cases for the general point about the need to relativize reliability conditions to methods (1980, p. 179 ff).

> whatever is their conventional meaning. Edmund has two further peculiarities. First, he is under the misapprehension that the English word "I" is a proper name that English speakers conventionally use to refer only to him. Second, he is in the pompous De Gaulle habit of using what he takes to be his own proper name to refer to himself. He intones self-importantly to his English audience "I will not tolerate this sniggering".

Edmund uses the word "I" with deferential semantic intentions. The word "I" is conventionally a device of self-reference in English. So Edmund self-refers using "I." He believes that he self-refers using "I" and he is in fact correct in believing this. However, his basis for believing that he is self-referring using "I" is the false lemma that "I" is his own proper name. Thus, he does not know that he is self-referring using "I." This is a Gettier case, for knowledge of linguistic self-reference.

Edmund is nevertheless linguistically expressing a first person thought, about what he himself will not tolerate. Therefore the Knowledge View is mistaken. The Intention View does better. Edmund, despite his semantic confusion, still speaks with the intention to self-refer.

The cases of Gertrude and Edmund make it plausible that the knowledge that one is self-referring is not necessary for the linguistic expression of first person thought. Is it nevertheless *sufficient* for the linguistic expression of first person thought that one should use language in the knowledge that one is thereby self-referring? Perhaps, this direction of the Knowledge biconditional is still plausible.

As the following kind of example brings out, not even this much is clear. The example serves to illustrate again the merits of the Intention View.

> Alf is writing up, from his notes, the minutes of a meeting that he attended along with others. He writes "Alf was present. Betty was present. Gary was present…". Alf knows that Alf is his name. Thus he knows that in writing "Alf was present" he is thereby self-referring. However, Alf records these minutes in a very cool-headed, detached, state of mind. He has at the time of writing no particular interest in the question of whether he himself was among the attendees. He simply expresses in writing the thought that Alf was present. Since he does know that he is Alf, he is in a position to draw the consequence that he himself was present. However, he need not do so, and given his present interests, he does not do so. It is not a first personal thought that he expresses but a third personal thought. Nonetheless he does, in detached amusement, note the semantic fact that he is referring to himself in writing "Alf was present".

The possibility of such a case appears to show that it is not sufficient for the linguistic expression of a first person thought that one should know that one is thereby self-referring.

It might be thought that the case threatens the Intention View no less than the Knowledge View. Does not Alf intend to self-refer using "Alf"? He is intentionally using a word, which he knows refers to him. If so, then the case shows that using language with the intention to self-refer is also not sufficient for the linguistic expression of self-conscious states of mind.

In fact this worry does not stand up to scrutiny. The element of “detachment” that makes it plausible that Alf does not express a first person thought also makes it plausible that Alf does not intend to self-refer. In order to see why, one should observe the distinction between what is intended and what is merely a foreseen side effect of doing as intended. For example, suppose that I intend to indulge my taste for fast food. I believe that I will thereby put on weight. Do I thereby intend to put on weight? No, putting on weight is a foreseen consequence of acting as intended; but it is not itself an aim I have in acting.

Is there a way of testing for this distinction? An intuitive test is to ask what would have happened under counterfactual changes in the agent’s beliefs. Suppose I had come to believe that indulging my taste for fast food would not in fact result in my putting on weight. Would I have sought alternative means of putting on weight? No. But if I had genuinely intended, and not only expected, to put on weight, then I would have sought alterative means upon learning that indulging my taste for fast good would not contribute to that in the end.

In the case as described, self-referring is known to Alf to be consequence of using the name “Alf.” The intuitive test makes it plausible that self-referring is not an aim Alf has in using the name “Alf.” Suppose, counterfactually, that while writing up the minutes Alf had come to believe that “Alf” was not his own name. Would he have sought alternative means of self-referring? No. Insofar as he is giving expression to a detached third personal thought about the participants he would go on to use “Alf” just as before, unmoved by the belief that he is no longer thereby self-referring.

Suppose, instead, that upon learning that his name was not “Alf” but, say, “Omar,” Alf would have chosen to write “Omar was present” instead. That counterfactual truth is evidence that it was a genuine aim of his, in his actual use of “Alf,” to self-refer. But it is equally evident that he was not really thinking of the participants of the meeting in a detached, third personal, state of mind. Why the interest in self-referring?

The examples of Gertrude, Edmund, and Alf, each demonstrate the superiority of the Intention View to the Knowledge View. The linguistic expression of first person states of mind seems to march in lockstep with the use of language with the intention to self-refer. On the other hand, the knowledge that one is self-referring in using language, while no doubt a typical accompaniment, is neither necessary nor sufficient for the linguistic expression of first person thought.

5 Linguistic Expression

It is time to consider some potential counterexamples to the Intention View. While ultimately ineffective, these cases will allow us to get clearer about the notion of linguistic expression as it features in the view.

> Bert is hiding in a foxhole. The enemy search party approaches the foxhole. Bert thinks the first person thought that he doesn't want to die. He has no intention to speak his mind, for he does not want to give away his position. However, he loses control of himself. Bert blurts out aloud "I don't want to die!"

Why might this be thought to be a counter-example to the Intention View? Suppose that Bert linguistically expresses his first personal state of mind when he produces the words "I don't want to die." Then the left-hand side of the biconditional Intention View is true. Now suppose that it is agreed that Bert does not intend to speak at all. A fortiori he does not intend to use words to refer to himself. Then, the right-hand side of the biconditional is false. The Intention View is mistaken to claim that one linguistically expresses a first person thought only if one uses language with the intention thereby to self-refer.

In order to see the way through this case, the defender of the Intention View should note first that the case is underdescribed in an important respect: The mechanism of speech production is not specified. Suppose that the case works as follows: The intense shame of entertaining such a cowardly thought as the thought that he does not want to die causes Bert to have a seizure. The neural mayhem randomly stimulates his motor cortex in such a way that he vocalizes the English sentence "I don't want to die."

In such a case, it is perfectly obvious that Bert does not act with the intention to self-refer.On the other hand, his noise-making is not an intentional action at all. To that extent, his noises cannot seriously be regarded as the linguistic *expression* of his state of mind. Of course, one could define the notion of "expression" so loosely that any causal effect of a state of mind counts as its expression. But in that sense there can be no interesting connection between kinds of mental state and their linguistic expression: Any subject could be wired up so that arbitrary states of mind cause arbitrary vocalizations. Claims like the Knowledge View and the Intention View are of interest only if linguistic expression is taken to be a kind of rational intentional activity. Events of speech and writing are events for which it is appropriate to ask what were the reasons, purposes, intentions of the agent's so acting.

So let it be supposed that Bert does not have a seizure but genuinely acts in using the sentence "I don't want to die." The case still appears to be troublesome for the Intention View, for then he *is* engaged in linguistic expression when ex hypothesi he did not intend to speak at all.

However, we need to distinguish prior intention from intention in acting. The case is one in which Bert had no immediately prior intention to self-refer. It does not follow that in speaking he has no intention to self-refer. If his spontaneous action is a rationalizable action at all, then it can sensibly be asked what were his intentions in acting, even if, as with other cases of sudden or spontaneous action, these intentions do not reflect any prior process of explicit deliberation on the part of the agent. Now, suppose Bert is asked during a reflective debriefing session after the event: Why did you use the word "I"—rather than the word "you," or the word "Hitler," or the word "sausages." His plausible answer, revealing of his intention in action, would be: “to speak of myself.” Why else would he have chosen the word “I” to express his thought?[10]

A different kind of potential counter-example to the Intention View involves, not unplanned outbursts of speech, but rather more calculated linguistic action. The case is another opportunity for clarification of the notion of linguistic expression.

> Winnie has beaten Louis in a two-player game of cards. A third party, interested in the outcome of this game, puts to Winnie the question: so who won? Winnie is a very modest person, and is averse to any explicit first personal bragging about herself. So she does not linguistically express her first personal thought that she herself won by means of the English words “I won”. Instead—it being common knowledge that Louis was her sole opponent—she chooses a more oblique strategy of answering the question, uttering the words “Louis lost”.

The potential difficulty for the Intention View is as follows. It might be thought that, in the circumstances, Winnie intentionally linguistically expresses the first person thought that she herself won the game. However, she does not use language with the intention to refer to herself. The only referring expression she chooses to use is the name “Louis,” with the intention of referring to Louis. Thus, one might conclude, the Intention View is wrong to claim that the linguistic expression of first person states of mind requires one to speak with the intention to self-refer.

The right response to this example is, again, to get clearer about the notion of linguistic expression. On an occasion on which a speaker uses words expressively, there may be any number of thoughts causally antecedent to the action. It is not plausible to say that the speaker gives voice to each one of these thoughts when he speaks. It is not even plausible to say that the speaker gives voice to all those thoughts that he expects an audience, in the circumstances, to be able to figure out that he thinks. The direct linguistic expression of a thought is only one such communicative strategy.

Can we say more about the identification of those thoughts that are genuinely expressed by a speaker on an occasion—as opposed to those insinuated or implicated by other communicative means? This is not the place for a proper treatment of this large question, but expectations on the part of the speaker, about what an audience would know on the basis of his utterance, must still be central. However, the relevant expectations here are in a certain sense purely semantic ones: One linguistically expresses the thought that *p* when one expects one’s audience to know that one speaks truly if and only if *p*—where the audience’s knowledge of this biconditional is expected to derive solely from what the audience would know about the (interpreted) words one uses.[11] The basic, and intuitive, idea is that one gives voice to those thoughts that one communicates by just exploiting the meaning of one’s words—as opposed to other contingencies about the situation in which one speaks.

[10] See Hornsby and Stanley (2005) for more evidence that even fast and spontaneous speech involves rational word choice.

[11] See Rumfitt (1995) for a compelling and detailed defense of roughly this approach.

How does this work in the present case? When Winnie produces the sentence "Louis has lost" she speaks in the expectation that her audience will know that she speaks truly if and only if she herself has won. However, the audience's knowledge of this truth condition is not expected to derive solely from their knowledge of the meaning of the name "Louis," the predicate "_ has lost," and their mode of combination. While she expects that they will know that she speaks truly if and only if she herself has won, this is only because she expects them to bring to bear some supplementary circumstantial information about the character of the game played, viz., that Louis has lost if and only if she herself has won. All that she can expect them to know solely on the basis of knowledge of the words she uses is that she has spoken truly if and only if Louis has lost. Thus, it is the thought that Louis has lost which is the thought she linguistically expresses. This is the thought she actually puts into words.

In order to linguistically express her first person thought, then, Winnie must overcome her modesty. She must speak in the expectation that her audience will know that she has spoken truly if and only if she herself has won, where she expects them to know this just on the basis of what is known about the semantic values of the words she chooses. And in order to do this, she will need to choose a word she thinks is commonly known to refer to her. She could, in the circumstances that everyone thinks that "Winnie" is her own name, speak a la De Gaulle and produce the sentence "Winnie has won." Or, more likely, she will just choose to use the first person pronoun "I," a word typically known in English conversation to be a device of self-reference. Either way, she can give voice to her first person thought only by choosing to use a word to refer to herself. So, just as the Intention View claims, the linguistic expression of first person thought requires speaking with the aim of self-referring.

I will close by considering a final potential counterexample to the Intention View. The right response to this example further illustrates the notion of linguistic expression, and provides an opportunity to clarify the conditions for intending to do something.

> Arthur is highly influenced by Schopenhauer. He is especially persuaded of the following pronouncement: "that the subject should become object for itself is the most monstrous contradiction ever thought of". Accordingly he develops some strange theoretical views about the pronoun "I". He believes that it is not a device whereby each speaker x may refer to x—for that would be a monstrous contradiction. Rather, each speaker x is a fusion of two proper parts, a body and a spirit, and when a speaker x uses "I", x refers not to x but to the spirit of which x has a form of introspective awareness. He has a correspondingly deviant view about the semantics of predicates such as "_ is making a mess". He believes that the predicate is true of an individual just in case the individual is the spiritual part of a person who is making a mess. Arthur, in all other respects, and outside his study, is perfectly well immersed in the ordinary communal use of the word "I" to give voice to first personal states of mind. For example when he realizes that he himself is making a mess, he will, without any second thoughts, say "I am making a mess".

The difficulty for the Intention View arises as follows. While it is questionable whether it is a condition of intending to ϕ that one must believe that one will *succeed* in ϕ-ing, the following weak belief constraint at least seems plausible: one intends to ϕ only if one does not believe that it is impossible for one to ϕ. Though we may idly wish for what we think is impossible, we cannot intend to do what we think is impossible.

Arthur believes, on the authority of Schopenhauer, that reflexive reference is impossible. According to the weak belief constraint, then, Arthur cannot intend to reflexively refer. Nonetheless, in virtue of his practical immersion in the ordinary use of "I," he does linguistically express his first person states of mind. So it is not necessary for the linguistic expression of first person states of mind that one use language with the intention to reflexively refer.

What is the right reaction to this case? It seems to me that the case calls not for the abandonment of the Intention View but for the refinement of the belief constraint upon intending. Let it be agreed that Arthur's theoretical opinions about the nature of first person reference are sincerely held. The right thing to say about the case is that in so far as Arthur, outside his study, is using the word "I" with the rest of the community to express first person thoughts, his theoretical opinions must in a certain sense take a back seat.

For let it be supposed that on an occasion when he thinks that he himself is making a mess, Arthur says "I am making a mess" while keeping firmly in the forefront of his mind his opinions about the impossibility of linguistic self-reference. Can he linguistically express his first person thoughts while consciously in the grip of this philosophical opinion?

Recall, for Arthur linguistically to express his thought that he himself is making a mess, he must produce an utterance in the expectation that the audience will know that he has spoken truly if and only if he himself is making a mess, where their knowledge derives solely from knowledge of his interpreted words. In order to determine whether this condition is met, we need to distinguish two variants of the case. In one case, Arthur believes his community shares his semantic opinions. In another case, Arthur believes that he is the only enlightened one; he thinks that the rest of the community mistakenly believes that "I" is a device whereby x may refer to x. In both cases, his present and lively opinion is that "I" does not, and cannot, refer to its speaker. It refers to the spiritual proper part of the speaker.

Take the first variant. Does Arthur, in using "I," meet the conditions for linguistically expressing his thought that he himself is making a mess? In this variant, what Arthur can expect his audience to know solely in virtue of their understanding of his words is that he speaks truly if and only if *this spirit* is part of a person who is making a mess. For, they share his deviant opinions about the semantics of the term "I" and the predicate "is making a mess." He may well expect them to be in a position to draw the further conclusion—perhaps on the basis of the supplementary metaphysical view that each person is able to demonstrate introspectively only his own spiritual part—that he speaks truly if and only if he himself is making a mess. But the thought he puts into words is a thought about his spirit. He does not express his first person thought, even if he would appear superficially to observers with our understanding of "I" to be someone expressing first person thoughts.

The second variant is less clear-cut. This is in effect a case in which Arthur does not expect his audience to know the real meaning of his words. He does expect them to believe that he speaks truly if and only if he himself is making a mess, but he expects them to believe this only because he expects them to have the beliefs—false beliefs, as he sees it—that the term "I" refers to the speaker, and the predicate "is making a mess" is true of individuals making a mess. But this feature of the case makes it implausible to say that he is genuinely putting his first person thought to them. As Arthur sees it, only linguistic incomprehension on the part of his audience leads them to believe that he speaks truly if and only if he himself is making a mess. But one cannot be said to express a thought to an audience if one speaks in the expectation that they do not know the meaning of one's words. In so far as Arthur, in his splendid isolation, genuinely expresses a thought, then he speaks either to himself, or to an imagined audience who would understand his words. In either case, the second variant collapses into the first variant of the case, the variant in which Arthur believes his audience to understand the meaning of his words. But in that variant, as explained, Arthur expresses a nonfirst personal thought, about his spirit.

The foregoing shows that if Arthur's opinions about the impossibility of self-referring with "I" are fully engaged when he speaks, then it is to say the least unclear whether we really have a case of the expression of first person thought. Of course, a more likely scenario for someone with Arthur's deviant philosophical beliefs will involve a character whose seminar-room opinions do not really impinge upon his psychology when he is immersed in the practice of ordinary speech. A rationalization of his habitual linguistic actions will ascribe to him, as much as to anyone else, the intentional choice of a word of reflexive reference when expressing first person states of mind.

This in turn shows, what should be clear anyway, that the weak belief condition on intention is at best an idealization. Perhaps it is really a psychological impossibility to form the intention to ϕ while at the same time clear-headedly endorsing the thought that ϕ-ing is impossible. However that may be, if an opinion about the impossibility of ϕ-ing is not engaged and occurrent at the moment of action, only manifesting itself in theoretical reflection, then there is no obstacle to acting with the intention to ϕ. Arthur is, at worst, mildly irrational in so far as he is caught up in the expression of first person thought in the usual way while at the same time retaining the philosophical opinion that reflexive reference is impossible.

So the case is not a counterexample to the claim, made by the Intention View, that the linguistic expression of first person thought requires the intention to self-refer. If Arthur's deviant opinions *are* operative at the time of speech, then he does not express first person thoughts even if he is superficially conformal with those who do express first person thoughts. If his opinions are *not* operative at the time of speech, then they are no obstacle to his intentionally self-referring.

I began with the question of how self-conscious states of mind are expressed in language. The simple view that such states of mind are exactly those expressed by uses of first personal language was shown to be a mistake. Of the two more promising accounts of the conditions for linguistic expression of self-conscious states of mind, the Intention View has emerged as superior. The Knowledge View, for all its eminent advocates, seems not to be correct.

Acknowledgments For comments and discussion, thanks to Ray Buchanan, Hans Kamp, Daniel Morgan, Lucy O'Brien, Lea Salje, and participants in the Taipei IEAS conference.

References

Anscombe, G. E. M. (1957). *Intention*. Oxford: Blackwell.
Anscombe, G. E. M. (1975). The first person. In *Collected philosophical papers: Vol. 2. Metaphysics and the philosophy of mind* (pp. 21–36). Oxford: Basil Blackwell.
Evans, G. (1982). *The varieties of reference*. Oxford: Clarendon Press.
Hornsby, J., & Stanley, J. (2005). Semantic knowledge and practical knowledge. *Proceedings of the Aristotelian Society, Supplementary Volumes, 79,* 107–145. http://onlinelibrary.wiley.com/doi/10.1111/j.0309-7013.2005.00129.x/abstract.
Martin, M. G. F. (2001). Out of the past: Episodic recall as retained acquaintance. In C. Hoerl & T. McCormack (Eds.), *Time and memory: Issue in philosophy and psychology* (pp. 257–284). Oxford: Clarendon Press.
Nozick, R. (1980). *Philosophical explanations*. Cambridge: Harvard.
O'Brien, L. (2007). *Self-knowing agents*. Oxford: Oxford University Press.
Peacocke, C. (2008). *Truly understood*. Oxford: Oxford University Press.
Romdenh-Romluc, K. (2008). First-person thought and the use of "I". *Synthese, 163,* 145–156.
Rumfitt, I. (1994). Frege's theory of predication: An elaboration and defense with some new applications. *Philosophical Review, 103*(4), 599–637.
Rumfitt, I. (1995). Truth-conditions and communication. *Mind, 104*(416), 827–862.
Setiya, K. (2003). Explaining action. *Philosophical Review, 112*(3), 339–393.
Velleman, J. D. (1989). *Practical reflection*. Princeton: Princeton University Press.

Personal and Sub-Personal: Overcoming Explanatory Apartheid

Hong Yu Wong

Abstract Cognitive neuroscientific research provides a rich source of findings that require philosophical reflection. The meeting of philosophy and neuroscience raises different questions, including that of how neuroscientific discoveries can impact philosophical accounts of mental phenomena. An influential answer to this question proceeds through distinguishing between a personal and a sub-personal level of explanation (Dennett, *Content and Consciousness,* 1969). The thought, very roughly, is that we can distinguish a personal level of explanation, which is the proper province of philosophy, and that neuroscientific explanations, however interesting, are to be confined to the sub-personal level. Such a move simultaneously allows us to recognize the contribution of neuroscience, and also to contain it, so that it does not challenge the explanatory ambitions of philosophy. I will examine two instances of this strategy from McDowell (*Philosophical Quarterly* 44(175):190–205, 1994) and Hornsby (*Philosophical Explorations, 3*(1): 6–24, 2000). They employ the distinction between personal and sub-personal levels of explanation to institute a kind of explanatory apartheid between the two levels. I argue that their arguments for explanatory apartheid fail. This allows us to see why the choice between isolationism and eliminativism is a false dilemma.

1 Introduction

The rise of cognitive neuroscience in the last 2 decades marks a turning point in our understanding of mental phenomena and their neural underpinnings. New analytic techniques have led to a cornucopia of information from the neurosciences in recent years. Cognitive neuroscientists are now asking questions about phenomena that have been the traditional province of philosophers: consciousness, self-consciousness, action, and rational choice, among other things. Human choice behavior has been shown to be subject in unexpected ways to how potential choices

H. Y. Wong (✉)
Werner Reichardt Centre for Integrative Neuroscience, University of Tübingen,
Tübingen, Germany
e-mail: hong-yu.wong@cin.uni-tuebingen.de

Tzu-Wei Hung (ed.), *Communicative Action,* DOI 10.1007/978-981-4585-84-2_6,

are framed (Kahneman et al. 1982; Kahneman 2011). The neural preparation for some voluntary movements appears to precede and predict conscious awareness of movement initiation, challenging the idea that conscious choice is what determines movement initiation (Libet et al. 1983; Libet 1985; Haggard and Eimer 1999; Soon et al. 2008). There is evidence that much of human action is automatic, and even intentional actions that are deliberate are subject to numerous nonconscious influences and biases (Bargh and Chartrand 1999; Moors and De Houwer 2006). These are just some results that have attracted the attention of philosophers and the wider public.

Two aspects of these results deserve mention. The first aspect is the challenge they appear to present to a classical philosophical conception of human beings as the exemplars of rationality and autonomy. Man has been seen as a creature that stands apart from animals by virtue of his powers of self-control, reasoning, and reflection. Some of these results appear to indicate that this classical conception of human beings is more fragile than we thought. The springs of our actions are less transparent to us and less under our control than we initially thought. The second aspect is what these results tell us about the neural and psychological underpinnings of these rational capacities. Given that the conditions under which we can exercise these capacities are more fragile than we expected, it would seem that a philosophical account of these capacities would need to take some measure of the empirical complexity of these phenomena. At times, the force of some of these results has been over dramatized, and the naïve conception of human powers over simplified. But, overall, it would seem that given the interest and wealth of new results about mental phenomena from the neurosciences, philosophy cannot simply brush these aside.

There is no question that future philosophical work on mental phenomena and their explanation will need to be highly sensitive to the precise parameters of these empirical findings. But it must not neglect the concepts, distinctions, frameworks, and questions that have been established by philosophers over the years. The meeting of philosophy and neuroscience raises both methodological and substantive questions. The methodological questions are: How are we to proceed in doing philosophy of mind in light of the advances in the neurosciences? Must we do neurophilosophy? These methodological questions tie into the substantive questions: How do the mechanisms that neuroscience has discovered impact the philosophical articulation of mental phenomena? Must a philosophical account of some mental phenomenon allude to the psychological and neuroscientific underpinnings of the phenomenon in question?

In this chapter, I want to consider an influential answer to the substantive questions that proceeds through distinguishing between a personal and a sub-personal level of explanation. The thought, very roughly, is that we can distinguish a personal level of explanation, which is the proper province of philosophy, and that neuroscientific explanations, however interesting, are to be confined to the sub-personal level. Such a move simultaneously allows us to recognize the contribution of neuroscience, and also to contain it, so that it does not challenge the explanatory ambitions of philosophy. I will examine two instances of this strategy from McDowell (1994) and Hornsby (2000). They employ the distinction between personal and sub-personal levels of explanation to institute a kind of explanatory apartheid between the two levels.

1.1 Personal and Sub-personal Distinguished

The distinction between the personal and sub-personal was first introduced by Dennett in *Content and Consciousness* (1969). There he distinguishes between "the explanatory level of people and their sensations and activities" and "the sub-personal level of brains and events in the nervous system." If we accept that examining the functioning of a system as a whole, as opposed to the functioning of some of its parts, introduces different levels of explanations, we can agree that we can distinguish between a personal as opposed to a sub-personal level of explanation. The former concerns people, and the mental states, events, and processes in their lives, which characterize them as a whole; the latter concern certain goings on in their nervous systems. In fact, we can think of the distinction between the personal and the sub-personal as a special case of that between the animal and the sub-animal (McDowell 1994). Having distinguished explanations concerning the whole, as opposed to explanations concerning parts of the whole, a further question concerns how these different levels of explanation relate.

In a critical passage, Dennett describes the chasm between these modes of explanation:

> When we have said that a person has a sensation of pain, locates it and is prompted to react in a certain way, we have said all there is to say within the scope of this vocabulary. We can demand further explanation of how a person happens to withdraw his hand from the hot stove, but we cannot demand further explanations of terms of 'mental processes'. Since the introduction of un-analyzable mental qualities leads to a premature end to explanation, we may decide that such introduction is wrong, and look for alternative modes of explanation. If we do this we must abandon the explanatory level of people and their sensations and activities and turn to the sub-personal level of brains and events in the nervous system.

So far, Dennett has distinguished between personal and sub-personal levels of explanation by observing that personal level explanations (in terms of "mental processes") run out somewhere—they come to an end—even though there may be explanatory demands that remain unsatisfied. This dissatisfaction is allayed by turning to explanations at the sub-personal level, in terms of the functioning of one's nervous system. However, in so introducing the distinction between the personal and the sub-personal, he also tells us about certain conditions on the descent to the sub-personal. The descent to the sub-personal satisfies a thirst for further explanation, when personal level explanations have come to an end—perhaps earlier than we might have liked. But this gratification comes at a cost: The recourse to sub-personal explanations comes with the abandonment of explanations at the personal level. He continues:

> But when we abandon the personal level in a very real sense we abandon the subject matter of pains as well. When we abandon mental process talk for physical process talk we cannot say that the mental process analysis of pain is wrong, for our alternative analysis cannot be an analysis of pain at all, but rather of something else—the motions of human bodies or the

> organization of the nervous system. Indeed, the mental process analysis of pain is correct. Pains are feelings, and felt by people, and they hurt….
>
> Abandoning the personal level of explanation is just that: abandoning the pains and not bringing them along to identify with some, physical event. The only sort of explanation in which pain belongs is non-mechanistic; hence no identification of pains or painful sensations with brain processes makes sense… (Dennett 1969, p. 105; emphases in original)

Dennett is making several points here. First, in moving to a sub-personal level of explanation we are changing the subject matter, the target of explanation. Second, both the personal level and the sub-personal levels of explanation are appropriate to their respective subject matters. Combined with the first claim that we have different subject matters at different levels, this gives us the claim that the levels of explanation are autonomous. Third, these levels of explanation are not only autonomous, but are isolated—since recourse to one requires abandoning the other. These three points may be labeled: *difference*, *autonomy*, and *isolation*.

Though introduced in the context of pain, and sensory awareness more broadly, these thoughts are meant to be equally applicable to action. Just as we can only talk of pains when we talk of people and their feelings, we can only talk of agency when we are in the realm of agents in the space of reasons (Hornsby 2000). So the claims that Dennett is making here in distinguishing between personal and sub-personal levels of explanation are general. There is no doubt more that could be said about Dennett's distinction (see, e.g., McDowell 1994; Davies 2000; Hornsby 2000). However, I want to probe the claim of isolation. In particular, I want to suggest that we should hold on to the claims of difference and autonomy, but that this does not require that we have to be isolationist. This would then open the possibility of a rapprochement between philosophy and neuroscience that is less protectionist. We need not ring fence the subject matter of philosophy in order to recognize the insights of neuroscience. But to do that we need to examine the arguments that proponents of this picture have put forward. Dennett (1969) merely hints at the defense of his distinction and its purported isolationist consequences. McDowell and Hornsby, however, take this forward, the former in the domain of perception and the latter in the domain of action. It is to these arguments that we now turn.

2 Arguments in Favor of Explanatory Apartheid

Dennett has since abandoned his earlier explanatory strictures without quite saying why. After *Content and Consciousness*, he has turned increasingly toward the cognitive sciences in an attempt to undermine certain classical philosophical conceptions of the human being, replacing these with philosophical projects with a neurophilosophical bent. Thus, his later work has become a target for philosophers who adhere to his earlier articulation of the distinction between the personal and sub-personal. The most prominent attempts to defend Dennett's early distinction are due to McDowell (1994) and Hornsby (2000). We shall look at each in turn.

2.1 McDowell Versus (Later) Dennett on Perceptual Phenomenology

McDowell (1994) takes the later Dennett to task for flouting his own explanatory strictures in propounding an account of perceptual consciousness (Dennett 1978).[1] The argument is largely negative. It takes as a starting point Dennett's description of visual phenomenology, and traces unsatisfactory aspects of this description back to Dennett's account of visual experiences as consisting in the operation of certain sub-personal mechanisms.

Describing his visual phenomenology in reading, Dennett writes:

> Right now it occurs to me that there are pages in front of me, a presentiment whose aetiology is not known directly by me, but which is, of course, perfectly obvious. It is my visual system that gives me this presentiment, along with a host of others. (Dennett 1978, p. 166)

McDowell rightly points out that as a description of ordinary visual phenomenology this is "off key." The ordinary phenomenology of seeing pages of a book is not one of having some intuitive feeling that a book is in front of me. This makes seeing sound like blindsight (Weiskrantz 1986), as if seeing were like a form of educated guessing.

If the issue were merely a slip in phenomenological description, this would not be of much interest. Phenomenology is difficult to do; and it is often hard to convey experiences in words. McDowell argues that we can partly trace Dennett's take on the phenomenology to his picture of how the personal level phenomenon of visual experience is to be explained. We get a glimpse of this in the second sentence of the quote above. In rough outline, Dennett's idea is that one's visual system provides one with "presentiments" of what one sees, based on its analysis of the ambient light. The idea appears initially plausible: Surely, vision requires the operation of one's visual system, and our visual experience is the upshot of the computations of one's visual system. No one should disagree with that. But Dennett's claim here is that one's visual system functions to deliver content to one. The radical nature of that claim can be brought out by comparison with a claim that Dennett endorses about frogs. In discussing a classic paper in neuroscience, "What the Frog's Eye Tells the Frog's Brain" (Lettvin et al. 1959), Dennett endorses a suggestion by the distinguished computational neuroscientist Michael Arbib that the slogan "What the frog's eye tells the frog" (p. 163) would better capture what goes on. The situation is supposed to be analogous for us: Vision consists in my visual system talking to *me*. Thus, Dennett thinks it is appropriate to speak of "my visual system [giving] me this presentiment [of what I see]."

McDowell points out that this cross level talk is confused. My seeing requires the proper operation of my visual system; some might even argue that my seeing just consists in the proper operation of my visual system. But just as the frog's eye

[1] McDowell (1994, p. 190, fn. 1) reports Dennett as being in agreement that what McDowell objects to in his 1978 article is not superseded by the account in his more recent book-length treatment of the topic, *Consciousness Explained* (1991).

does not talk to the frog, our visual systems do not talk to us. The frog's eye talks to the frog's brain. My seeing does not consist in the visual system delivering some message to me. Rather, my visual system's proper operation enables me to have the visual experience I have of the world.

The lesson we are to take from this is that we have to distinguish between the personal and sub-personal levels of explanation on pain of confusion. The argument here has the character of pointing out a category mistake. Dennett's mistake is to take the sub-personal mechanisms that enable visual experience for a constitutive condition. No doubt the information processing involved in the operations of the visual system is complex and interesting. But these are enabling conditions on vision and visual experience. However critical these are, they are not to be confused with constitutive conditions on visual experience; taking them as constitutive conditions, like Dennett, only leads to an incoherent picture. I shall not attempt to work out the full details of how McDowell's argument proceeds or whether it is ultimately successful.[2] But the lesson is clear: There is no space for cross level explanatory claims.

2.2 Hornsby Versus Kim on Action and Mental Causation

Hornsby's (2000) argument concerns agency and its exercise. Her target is the entire debate on mental causation due to Kim's causal exclusion argument (Kim 1998). The issue Kim presses is this: Physicalists who believe in the nonidentity of mental and physical events face the very same problems as substance dualists about the causal efficacy of the mental, since, for whatever effect the mental event has, there is always a sufficient physical cause for that same effect. Hornsby thinks that this entire debate is confused, because even though the considerations that motivate the problem begin with action—how can my decisions, plans, and intentions bring about my bodily movements?—we no longer have *action* in view once we are in the realm of brain states.

Her argument takes the form of a reductio ad absurdum which relies on the notion of teleologically basic actions (Hornsby 1980). Roughly, teleologically basic actions are those actions the performance of which does not require procedural knowledge of how to perform any other action. When an action is not done by exercising procedural knowledge of how to perform some other action, we can say that that action is performed directly. For example, once one learns how to tie one's shoelaces, and does so as a matter of habit, in exercising one's knowledge of how to tie one's shoelaces, one does not require knowledge of how to tie one's shoelaces by first making a loop, then another, crossing over, and so on. Another familiar

[2] My discussion of McDowell's argument considers only aspects which immediately relate to the distinction between the personal and the subpersonal, and between constitutive and enabling conditions. McDowell also objects to other aspects of Dennett's account, such as whether perceptual content can be singular, his perceptual epistemology, and his discussion of nonconceptual content, among other things. McDowell attempts to weave these issues all together; my discussion distils McDowell's key point concerning Dennett's mistaken crossing of levels.

example is playing scales on a musical instrument, say, the violin. First, one does it step by step: While drawing the bow, one learns to put one finger down to play the first note, and then another to play the next note, and so on, until one has played an octave. But once one has learnt to do that and has the skill to do so, one can simply play the scale without exercising procedural knowledge how to do each step and then the other and so on. Many of our ordinary bodily movements have this character—we do not do them through exercising knowledge of how to do something else.

Once we have the notion of teleologically basic action in view, we can proceed to state Hornsby's argument. On Descartes's view, the only organ that one can move is the pineal gland. But souls do not think about moving the pineal gland at all, and have no procedural knowledge how moving that moves one's bodily extremities. But any bodily action would require moving one's bodily effectors by moving the pineal gland in some way. Since we do not appear to have any knowledge of how to move one's body by moving one's pineal gland in some way, action does not appear to be possible.

Quite surprisingly, contemporary physicalism is subject to a similar problem, Hornsby argues. Consider an action of raising one's hand. This is a simple action that is teleologically basic. Kim's problem is that brain states screen off the mental states, like intention, which supervene on them. Thus, we may ask whether the brain moves the hand directly or indirectly. If the brain is able to move one's hand directly, this would be mystifying, as there is quite some distance between the brain and one's hand. So it must be done indirectly. However, if the brain has to move the hand indirectly, it would have to know how to move the hand by moving something directly. But the brain does not have such knowledge. The upshot of this is that brains are not the kinds of things that can move hands. Thus, the causal exclusion problem does not arise.

The lesson we are to learn from this dialectical episode is that to keep action in view, we cannot leave the personal level. The problem is not that brain states joust for causal influence with one's mental states, but that to understand how people can act, what we need is not a descent to sub-personal machinery, but an understanding of the agentive capacities people can exercise. The thought, similar to Dennett's—when he was introducing the distinction between the personal and the sub-personal—is that personal level explanation comes to an end sooner than expected, but that is all we have: "As Ryle and Wittgenstein might have said, the correct answer to the question 'How can she move her hand when she wants to?' or 'How does she move her hand?' is 'She just can'" (Hornsby 2000, p. 14). Again this is not to deny that the neural systems for motor and executive control are involved in the control of movement, and that there is much complexity and interest in the neural systems controlling movement. But Hornsby's thought is that alluding to these neural systems is changing the subject. A properly functioning motor system is an enabling condition on bodily action, not a constitutive condition. Once again, we are mired up in confusion if we cross from the realm of personal explanation into the sub-personal for illumination.

To conclude: The picture McDowell and Hornsby present, expanding on the ideas of early Dennett, is clear. They intend to institute an explanatory apartheid between the personal and sub-personal levels of explanation. We may express this as a slogan: *Personal is Personal, and Sub-Personal is Sub-Personal, never the twain shall meet…*[3]

3 Reflecting on the Arguments for Explanatory Apartheid

There is something correct in McDowell and Hornsby's protests. Their insistence that a philosophical account of the person requires that we keep the personal level in view is to be lauded. If we seek an explanatory project of elucidating personal level capacities just in terms of sub-personal machinery, we risk losing sight of the personal level. Furthermore, they show that in articulating a philosophical account of a person's capacities, alluding to sub-personal explanations does not always help. This is all excellent. But their arguments attempt to establish something much stronger. They want to show that there is no space for cross level explanatory claims of a certain sort: In particular, sub-personal explanations cannot figure in a constitutive explanation of personal level capacities except as enabling conditions.

I do not think that McDowell and Hornsby manage to establish this. Such a general claim would require showing that any attempt at cross level explanatory claims would be flawed. The shape of their argument has a different character. They each take one prominent project, and argue that it is confused because there is an illicit crossing of levels. Even if we agree that McDowell and Hornsby's criticisms are on the mark against their respective targets, this does not yet establish their general claim that there is no space for cross level claims. What is required is the further assumption either that (1) the views that they criticize are representative, and, in particular, that the mistakes made are somehow paradigmatic, so that the possibility of cross level claims stand or fall with them, or that (2) any cross level claim will lapse into the kind of category mistake or confusion that they identify.

Neither assumption is without controversy. Dennett's view that the visual systems delivers presentiments to the subject is rather peculiar, and many philosophers who are attracted to drawing on sub-personal explanations in elucidating perceptual experience would shy away from Dennett's picture (e.g., Campbell 2002; Burge 2010). So Dennett's view cannot be said to be representative. There is more to be said on this front for the problem of mental causation. This is seen as a critical problem for physicalism that calls for resolution. However, even if this problem dissolves once we recognize that human beings have the capacity to act, it does not follow that no cross level claims in the domain of action are possible. It has not been shown that any cross level claim will lapse into category mistake or confusion. Without a formula to show that any such claim will fail, no grounds have yet been given for the impossibility of cross level claims.

[3] Compare Rudyard Kipling: "Oh, East is East, and West is West, and never the twain shall meet." *Barrack-room Ballads* (1892).

Are there ways to push for a chasm between personal and sub-personal levels of explanation based on reflecting on the concepts of the personal and the sub-personal? One way is perhaps to define a proper domain of philosophy. Philosophy is to be seen as the guardian of the personal level, and the task of philosophy is to articulate personal level claims and explain these. The observation that philosophy uses primarily armchair methods gives this claim some plausibility. Even if philosophy is seen in this light, it is unclear that explaining personal level claims should only draw on resources from within the personal level. Nothing about the concept of the personal demands this.

Furthermore, it is unclear that there are limits on the proper subject matter of philosophy. We can have philosophical claims about the sub-personal just as we can have philosophical claims about the personal level. The converse is also false. The personal level does not appear to be the sole domain of philosophy. It would be incorrect to say that neuroscience has nothing to say about the personal level. On Marr's (1982) influential demarcation of levels of explanation in vision research, he distinguished between the computational, algorithmic, and implementation levels. The computational level specifies the problem to be solved, the algorithmic level specifics the algorithm solving the problem, and the implementation level specifies how the algorithm solving the computational problem is implemented. The claims that Marr considers at the computational level are not unlike what McDowell considers at the personal level, such as what the function of vision is in a creature's life. So a defense of explanatory apartheid through isolating the subject matter of philosophy will not work.

Yet another attempt is to derive explanatory apartheid through mapping personal level explanations to constitutive explanations and sub-personal level explanations to enabling explanations. This is consonant with McDowell's picture. But the strategy relies on our being able to explicate the notions of constitutive as opposed to enabling conditions, and constitutive as opposed to enabling explanations of some particular phenomenon. The intuitive contrast between the two notions is clear, yet it is notoriously difficult to explicate the notions. Most philosophers, including McDowell, simply employ the contrast without explicating it in any way (but see Burge 2010, Chaps. 2 and 11 for some discussion). The idea, very roughly, is that constitutive conditions of something elucidate the nature of that thing. By contrast, enabling conditions do not, but are background conditions required for its existence. Correspondingly, constitutive explanations are those explanations based on the constitutive conditions that capture the nature of things. For example, many philosophers of action would be tempted by the following picture. On the standard story of action (e.g., Davidson 1980; Bratman 1987), actions are those events that are appropriately caused and rationalized by some select motivational antecedents, the most familiar of which is intention. We could say that on orthodox philosophical views of action, intention is a constitutive condition on intentional bodily action, but the functioning of one's motor system is an enabling condition. Another example comes from McDowell's discussion of visual experience: Perception constitutively is a creature having a bit of the world in view; this is enabled by the complex operation of its visual system.

There is no doubt that more must be said about the contrast between the notions of constitutive and enabling conditions, and the corresponding notions of explanation. However, we have the materials to rule out certain claims as deriving from the very notions of constitutive and enabling conditions. An initial temptation might be to equate the notion of personal explanation with that of constitutive explanation and sub-personal explanation with enabling explanations. But this cannot be correct, since there can be constitutive explanations of sub-personal, and indeed, non-personal phenomena. Yet the point that the proponents of explanatory apartheid wish to push is subtler. Their thought is that the constitutive explanations of personal level phenomena, phenomena that concern the person as a whole, cannot allude to sub-personal mechanisms. As we have observed, nothing in McDowell or Hornsby's arguments demonstrate that this is *generally* incoherent or impossible. But in considering the notion of constitutive conditions and the corresponding notion of constitutive explanation, it also does not appear to demand that personal level phenomena can only be elucidated by personal level phenomena or observations. Of course, it is correct to say that enabling conditions for some phenomenon will not provide a constitutive explanation of that very phenomenon. But this does not show that no constitutive explanation of personal phenomena can allude to sub-personal mechanisms.

Constitutive explanations simply require whatever it takes to explain the nature of the target of explanation. But nothing in the concept of a person or personal level phenomena show that their nature must be elucidated only by notions at that level. This is not to say that the nature of persons can receive elucidation by sub-personal notions, but only to point out that the notions of personal, sub-personal, and that of constitutive and enabling explanations do not yet prohibit it. Dialectically, we are back to where we started with McDowell and Hornsby. They do not have any general argument against the possibility of sub-personal mechanisms playing some role in constitutive explanations of personal level phenomenon. In the absence of further argument, then, there is no principled barrier to cross level explanations. There are, no doubt, many bad and confused cross level explanations, but this is no support for explanatory apartheid.

4 A False Dilemma: Isolationism or Eliminativism

Our reflections suggest that the defensive measures precipitated by the meeting of philosophy and neuroscience results from a false dilemma: isolationism or eliminativism. The fear is that as philosophers of mind faced with the breakthroughs of neuroscience, we have to turn either to some form of isolationism or some form of eliminativism. The latter is self-defeating and thus unacceptable. A weaker response is to think that we must do philosophy of mind by doing what is nowadays called "neurophilosophy." (This is a neurophilosophy that is shorn of the original eliminative commitments of the Churchlands). Some of what is done under the banner of neurophilosophy is more or less traditional philosophy of mind, but there is a clear movement toward the idea that philosophical theses have to be subject to empirical verification or falsification. This requires that they must be in some way operationalizable so that they can be empirically tested. While we should not deny that this way of doing philosophy of mind may throw up interesting ideas, there is no clear reason why this would be the only way of doing philosophy in the face of neuroscience.

The other choice that we initially appear to be pressed toward is isolationism. This intellectual pressure is nicely expressed by Hornsby when she writes that: "when realism in philosophy of mind is a doctrine about a *sui generis* personal level, scientific findings cease to threaten it. The personal/sub-personal distinction can protect one against eliminativism" (Hornsby 2000, p. 22). While we may feel the threat of another discipline encroaching on the traditional subject matter of philosophy, there is no need—and, I think, there is no way—to isolate the proper domain of philosophy. Thus, these two horns constitute a false dilemma. We should not succumb to the temptations of either horn in an episode of existential crisis.

5 Conclusion: Looking Forward

In this chapter, I have shown that certain prominent arguments for explanatory apartheid based on drawing a sharp distinction between personal and sub-personal levels of explanation does not have the general force that has been claimed for them. While we have shown that there are no general grounds for the impossibility of sub-personal mechanisms playing some role in constitutive explanations of personal level phenomena, we have not shown that there are such explanations.

I want to close by briefly considering two strategies forward in pushing for sub-personal mechanisms as potentially explanatory of personal level phenomena. The first is to attack the distinction between constitutive and enabling conditions as mutually exclusive and exhaustive. It is unclear that if some conditions do not fit the mould of paradigm examples of constitutive conditions, then they must be enabling conditions. Rather, sub-personal mechanisms might play a key role in an illuminating philosophical account of some personal level phenomena, without falling into the category of either constitutive or enabling conditions. Some of Campbell's work on perceptual consciousness can be seen in this light (Campbell 2002). A second strategy is to attempt to show that some constitutive explanations of a specific personal level phenomenon must draw on the articulation of the relevant sub-personal mechanisms. Burge has attempted such a project for perception (2010), and Butterfill and Sinigaglia (2014) have argued that certain motor representations are as good candidates for providing constitutive conditions for action as intentions.

Ultimately, philosophy is concerned to give the most general account of the nature of things; and there is no saying ahead of time what things will enter into the explanations that will be needed for understanding the personal level as such. In articulating the picture of mind for our time, we should delight in being able to draw both on the rich tradition of philosophy and the new discoveries that the neurosciences are throwing up. This way we can simultaneously recognize the pleasures of cognitive neuroscience and the consolation of philosophy.

Acknowledgments I am indebted to Tzu-Wei Hung for his invitation to the 2013 IEAS Conference on Language and Action at Academia Sinica, which spurred these reflections. Some of this material was also presented in seminars at Cologne and Tübingen. I am grateful to the audiences on all these occasions for their feedback. I wish to thank Rüdiger Bittner, Jennifer Hornsby, Johannes Roessler, and especially Krisztina Orbán for instructive discussions on these topics.

References

Bargh, J., & Chartrand, T. (1999). The unbearable automaticity of being. *American Psychologist, 54,* 462–479.

Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge: Harvard University Press.

Burge, T. (2010). *Origins of objectivity*. Oxford: Oxford University Press.

Butterfill, S. A., & Sinigaglia, C. (2014). Intention and motor representation in purposive action. *Philosophy and Phenomenological Research, 88*(1), 119–145.

Campbell, J. (2002). *Reference and consciousness*. Oxford: Oxford University Press.

Davidson, D. (1980). *Essays on actions and events*. Oxford: Oxford University Press.

Davies, M. (2000). Persons and their underpinnings. *Philosophical Explorations, 3,* 43–62.

Dennett, D. (1969). *Content and consciousness*. New York: Humanities Press.

Dennett, D. (1978). Toward a cognitive theory of consciousness. In D. Dennett (Ed.), *Brainstorms: Philosophical essays on mind and psychology* (pp. 149–173). Cambridge: The MIT Press, A Bradford Book.

Dennett, D. (1991). *Consciousness explained*. Boston: Little, Brown and Company.

Haggard, P., & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research, 126*(1), 128–133.

Hornsby, J. (1980). *Actions*. London: Routledge and Kegan Paul.

Hornsby, J. (2000). Personal and Sub-Personal: A defence of Dennett's early distinction. *Philosophical Explorations, 3*(1), 6–24.

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Kim, J. (1998). *Mind in a physical world*. Cambridge: MIT Press.

Kipling, R. (1892). *Ballads and barrack-room ballads*. New York: Macmillan.

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Journal of the Institute of Radio Engineers, 47,* 1940–1951.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences, 8,* 529–539.

Libet, B., Gleason, C., Wright, E., & Pearl, D. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act. *Brain, 106,* 623–642.

Marr, D. (1982). *Vision*. New York: W. H. Freeman.

McDowell, J. (1994). The content of perceptual experience. *Philosophical Quarterly, 44*(175), 190–205.

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin, 132,* 297–326.

Soon, C. S., Brass, M., Heinze, H. J., & Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nature neuroscience, 11,* 543–545.

Weiskrantz, L. (1986). *Blindsight: A case study and implications*. Oxford: Oxford University Press.

Part III
Sensorimotor Interaction and Language Acquisition

How Infants Learn Word Meanings and Propositional Attitudes: A Neural Network Model

Alistair Knott

Abstract An influential proposal by the developmental psychologist Michael Tomasello is that infants only properly begin learning word meanings when they acquire the concept of a *communicative action*, which happens around the age of 12 months. While Tomasello advances interesting empirical evidence for this proposal, he does not make any suggestions about how communicative actions are represented in the infant brain, or about the mechanism through which an understanding of communicative actions facilitates word learning. In this chapter, I will present a neural network model of language and cognitive development which addresses both of these questions. The representations of communicative actions that the model learns (which have roughly the form *X says that P*) encode the propositional content of utterances in a novel way. I also discuss how these representations may serve as developmental precursors for more sophisticated propositional attitude representations such as *X believes that P*.

1 Tomasello's Account of Word Learning and Pragmatic Development

Michael Tomasello's influential model of language development (Tomasello 2000; Tomasello 2003) emphasizes the role of infants' pragmatic understanding of the world in supporting their learning of language. For all humans, whether adult or child, language has a pragmatic function: We communicate linguistically in order to further social goals (for instance to share our beliefs and desires with others, or to ascertain the beliefs and desires of others) or to pursue joint undertakings (for instance to collaborate in a shared task). Tomasello imagines an infant observing a speaker producing an utterance directed at a hearer (possibly the infant herself). He proposes that the infant's understanding of the speaker's intentions in this context plays a crucial role in her ability to learn an association between the linguistic form of the utterance and its meaning. The infant interprets the meaning of the utterance in the light of her understanding of the speaker's current goals, and of how the hear-

A. Knott (✉)
Department of Computer Science, University of Otago, Dunedin, New Zealand
e-mail: alik@cs.otago.ac.nz

Tzu-Wei Hung (ed.), *Communicative Action,* DOI 10.1007/978-981-4585-84-2_7,

er features in these goals. At some point during development, the infant becomes aware of the general fact that human agents perform actions in service of goals. In Tomasello's account, learning this fact is a precondition for learning language. When the pragmatically aware infant observes a speaker producing an utterance, she attempts to infer the speaker's goals, and then forms hypotheses about how the words in the utterance further these goals.

In Tomasello's model, an early instance of pragmatic learning in language development concerns the learning of the meanings of individual words. Words are symbols that denote concepts. While some theories of language development assume the relationship between words and concepts is founded simply in the existence of regular associations or co-occurrences between words and concepts in the infant's mind, for Tomasello it is fundamentally pragmatic in origin, and has to be learned by infants in pragmatic analyses of speech events. What infants must learn is that uttering words can serve to evoke representations in the mind of the hearer. It is only after learning this general fact that infants can properly begin to learn the meanings of particular words.

Tomasello sees two pragmatic abilities as prerequisites for learning word meanings. One is the ability to establish *joint attention* with an observed agent, i.e., to attend to the same object the observed agent is attending to. When an infant has acquired this ability, a speaker's gaze will direct her attention to particular objects. Since speakers often visually attend to the situations they describe linguistically, the infant can learn that words can likewise serve to direct attention to arbitrary concepts. The other important pragmatic ability is the ability to infer the *communicative intentions* of an observed agent, i.e., to infer the goal underlying the agent's communicative actions. The infant must learn that there is a special class of actions (communicative actions) which have communicative effects rather than physical effects. Communicative actions are physical actions, which are directed at another agent, who is physically present in the communicative situation. But unlike regular physical actions, their effect is on the agent's mental state rather than on his physical state: Specifically, they evoke representations in the agent's mind. In spoken language, these actions are articulatory gestures that realize phonological word forms. Tomasello argues that the infant must be able to identify the special communicative effects of such actions before she can learn to associate specific actions with specific effects.

Tomasello advances both conceptual and empirical arguments for his proposal about word learning. The conceptual arguments turn on the question of what it means for words to be symbols: As just summarized, Tomasello argues that the meanings of words are more than just concepts which are regularly associated with them. The empirical arguments are of three types. Firstly, he argues that infants acquire the social-pragmatic skills needed to learn words (joint attention and the ability to recognize the intentions underlying communicative actions) around the age of 9–12 months—and that this is also the age at which infants start to learn word meanings (see, e.g., Tomasello 1995). Secondly, he argues that these social-pragmatic skills mark a key difference between humans and their closest evolutionary cousins, great apes (see, e.g., Tomasello and Herrmann 2010). Finally, he cites evidence that human infants systematically interpret speakers' utterances with reference to their inferred intentions (see, e.g., Diesendruck et al. 2004).

While there are good grounds for Tomasello's proposal about the role of social-pragmatic abilities in learning word meanings, the exact nature of these abilities is less clear. For one thing, it is unclear how an infant *represents* the communicative intentions of a speaker, or the special properties of the actions which achieve communicative intentions. We do not have good models of how intentions of any kind are represented in the brain. There are promising models of the cognitive states which store prepared physical actions (e.g., Miller and Cohen 2001) or prepared sequences of physical actions (e.g., Averbeck et al. 2002), but these are some way from providing a full model of the outcomes which agents intend when they perform actions. And the representation of *communicative* intentions presents particular problems. A communicative intention is an intention to bring about a certain mental state in the mind of a hearer—for instance, to make the hearer entertain a certain belief about the world. We do not have good models of how agents represent the current beliefs of other agents, let alone their intended beliefs. In Tomasello's model, infants draw inferences about the communicative intentions of speakers, and use these (along with the ability to establish joint attention) to help learn word meanings—but Tomasello does not make any suggestions about how these inferred intentions are represented in the infant brain. Moreover, he does not make any suggestions about how the development of social-pragmatic abilities in infants leads to their ability to learn word meanings. Tomasello's account proposes that infants' social-pragmatic understanding influences their ability to learn word meanings—but he does not give an account of the neural mechanisms through which this influence is exerted.

On the one hand, it is prudent to express the social-pragmatic theory of language learning at a high level, given that we know so little about the neural representations of intentions and mental states. On the other hand, if our interest is in understanding how the brain represents intentions and mental states, Tomasello's developmental account provides potentially useful information. Presumably infants' earliest representations of the mental states and communicative intentions of other agents are fairly simple, and become more complex as development progresses. It may be easier to model early protomental states and protocommunicative intentions than to model the mature representations that eventually develop—and a model of the representations which emerge early in development may contain clues about the mature representations that emerge later.

In this chapter, I will introduce a neural network model that addresses both how infants represent the communicative actions of observed agents and how this representation supports the learning of individual word meanings. I will begin in Sect. 2 by reviewing some of the difficulties to be tackled in formulating a model of communicative action representations. In Sect. 3, I will present a neural network model of vocabulary learning and its interaction with the development of a simple concept of communicative actions. In Sect. 4, I will discuss how this model suggests some solutions to the problems inherent in representing communicative actions, and I will conclude with a discussion of the model in Sect. 5.

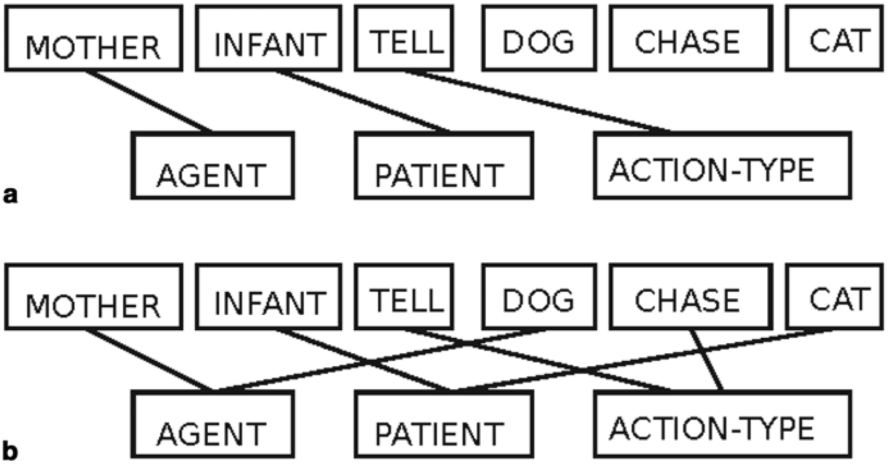

Fig. 1 **a** Representation of a single proposition by association of concepts with roles. **b** Problematic representation of nested propositions

2 Neural Representations of Communicative Actions

Consider a situation in which a mother tells her infant that a dog is chasing a cat. In Tomasello's model, the infant represents this action of her mother in a way that highlights its special communicative status. What might this representation look like? I will discuss two issues that need to be addressed, that both relate to the fact that communicative actions *express propositions*: The mother tells her infant (that P is the case).

The first issue is specific to models of neural representation. Communicative action representations feature *nested* propositions. The outer proposition is about the act of communication (the mother tells the infant something); in addition, the material that is communicated is also a proposition in its own right (a dog is chasing a cat). Nested propositions are hard to represent in a neural network. Representing a simple proposition involves activating a collection of concepts (for instance, MOTHER, INFANT, TELL), but it is also important to indicate the roles these concepts play in the depicted action: For instance, TELL indicates the ACTION-TYPE, MOTHER is the AGENT of the action, and INFANT is its TARGET or PATIENT. There are several ways of binding concepts to roles. Most straightforwardly, we could create direct associations between concepts and role labels (Chang 2002), as shown in Fig. 1a. This works well for simple propositions, but it is problematic if there is a nested proposition. The roles played by concepts in the nested proposition must also be represented; in our example, where the communicated proposition is that a dog is chasing a cat, we must associate DOG with AGENT, CAT with PATIENT, and CHASE with the ACTION-TYPE. Simply overlaying these associations on top of those defining the outer proposition (as shown in Fig. 1b) creates problems: There is nothing in this representation to indicate *which* agent and patient participate in *which* action. Many solutions to this problem have been devised (see, e.g., Plate 2003; van der Velde and de Kamps 2006). For the moment, I just want to note that nested propositions pose special problems for neural networks, which require special solutions.

The second issue to be addressed concerns how the relation between the outer and inner propositions should be encoded. The outer proposition describes the act of expressing, or representing, the inner proposition. The verb *tell* links a person (the speaker) to a represented state of affairs in the world: In philosophical parlance, it expresses a *propositional attitude* of the speaker towards a certain state of affairs. An agent can adopt different attitudes towards a represented proposition: He can believe it, desire it, doubt it, fear it—he can also *tell* it to someone else. In each case, the agent's attitude is "about" a proposition: To use philosophical terminology once again, it has *intentionality* (Brentano 1874; for discussion see Dennett 1989; Jacquette 2004). This relation of "aboutness" between an agent's attitude and its propositional content is notoriously hard to define. But there are some well-known properties of statements about propositional attitudes that any account of this relation must capture. Firstly, when we assert that agent *A* adopts a given attitude towards proposition *P*, we commit ourselves to the existence of agent *A* and her adoption of this attitude, but we are not committed to the truth of *P*. For example, when I say that *Jane believes a dog is chasing a cat*, I am asserting that Jane exists, and has a certain belief—but I am *not* asserting the content of this belief, i.e., that a dog is chasing a cat. (I am not even committed to the *existence* of the dog and the cat. I can assert that Jane believes a unicorn is chasing a dragon without believing in unicorns and dragons myself.) Secondly, assertions about propositional attitudes are *intensional*: That is to say, their truth depends on the way the content of these attitudes is reported linguistically. For instance, assume the cat featuring in Jane's propositional attitude happens to belong to the prime minister, but that Jane does not know this. *Jane believes a dog is chasing a cat* is a true statement, but *Jane believes a dog is chasing the prime minister's cat* is not. In regular assertions about the physical world, the truth of a statement is not dependent on language in this way. For instance, if Jane is the prime minister's daughter, and Jane sneezed, then we can truly assert *The prime minister's daughter sneezed*: Whether Jane knows she is the prime minister's daughter or not is irrelevant. To account for these properties of propositional attitudes, logicians traditionally adopted modal logic as a knowledge representation formalism, allowing reference to possible worlds other than the actual world in representations of propositional content, and in representations of the meaning of words (see, e.g., Montague 1974). A more recent strategy is to model propositions and words as the cognitive states of agents, which may or may not reflect the current state of the world (see, e.g., Gärdenfors 2004): this is the approach I will take.

Statements about an agent's communicative actions express propositional attitudes in both the respects just described. The statement *X* tells *Y* that *P* asserts the existence of the speaker *X* and the hearer *Y* , and the fact that a telling event occurred, but it does not assert that *P*. And clearly, the words that report the telling action have a bearing on the action which is reported: asserting that *Jane told her* *daughter the dog was chasing the cat* is different from asserting that *Jane told her daughter the dog was chasing the prime minister's cat*. At the same time, telling (and other communicative actions) are unusual as propositional attitudes. Whereas the prototypical attitudes (believing, desiring, etc.) are pure mental states, communicative actions are substantive actions: they have motor components as well as mental components. Identifying a communicative action involves processing a speaker's physical actions. For a linguistic action, the action is typically an articulatory gesture expressing a sequence of word forms. Word forms are associated with concepts, and the manner in which word forms are assembled conveys information about how these concepts are connected to form propositions, so the gestures which form the physical component of a communicative action collectively convey the propositional content of the action. A hearer who knows the language being used can recover the propositional content of the action from the gestures. Identifying the "purely mental" propositional attitudes of an agent (e.g., beliefs and desires) is not so closely tied to the processing of a particular type of action. Mental states like desires and beliefs can be inferred from a variety of sources: for instance facial expressions or overt behavior. Of course they can also be inferred from linguistic utterances: if someone says *P*, a strong default is to infer they believe that *P*. But communicative actions are unique in being *tied* to particular movements: they convey propositional attitudes conventionally through physical movements.

This close connection with physical movements makes the development of communicative action representations a natural first step in the development of propositional attitude representations. Tomasello's proposal that infants must learn to identify communicative actions at an early point during language learning thus fits well within an account of the development of mature propositional attitude representations. But as discussed in Sect. 1, Tomasello does not say anything about how infants represent communicative actions.

In this section, I have outlined two requirements for any model of the communicative action representations developed by infants. Firstly, it must provide a means for representing nested propositions. Secondly, it must capture the intentionality of communicative actions: the elusive relation between a physical speaking action and the propositional content it expresses. I turn now to a computational implementation of Tomasello's developmental theory that aims to address the open questions about representations that it raises.

3 A Neural Network Model of the Role of Communicative Action Concepts in Word Learning

In this section, I will describe a neural network model of infant word learning which also gives an account of how infants develop simple representations of communicative actions. The model is intended to describe developmental processes occurring between the ages of around 10 and 18 months. In the model, infants' learning of word meanings and their development of communicative action representations bootstrap

Table 1 Input data to the network: parallel streams of concepts and word forms

Time		t_1	t_2	t_3	t_4	t_5
Concepts	AGENT	DADDY	COOKIE	DOG	MUMMY	CAT
	ACTION-TYPE	LAUGH		JUMP	TALK	RUN
Word forms		*puppy*	*supper*	*toy*		*cat*
		ball	*ready*	*break*		*run*

one another: learning word meanings facilitates the development of communicative action representations, which in turn facilitate the learning of word meanings. The model is called *pragmatic bootstrapping*; the network which implements the model is described in detail in Caza and Knott (2012). In the current chapter, my focus is on the representations of communicative actions which the network learns; technical details of the model can be found in Caza and Knott (2012).

3.1 Input Data

The input data for the network is a stream of word-form representations, and a parallel stream of conceptual representations. These simulate the inputs being received by an infant. The conceptual representations are assumed to be delivered by the sensorimotor system; these reflect the infant's real-time visual, motor and auditory experience of the world. The word-form representations are phonological encodings of words produced by mature speakers in the infant's current environment, again in real time. We assume that the infant is already able to identify individual words as phonological units, using the statistical learning abilities documented by Saffran et al. (1996). We also assume the infant is able to follow the gaze of an observed human agent—an ability which is also attested at 10 months, and which develops rapidly from 10 to 18 months (Butterworth and Jarrett 1991). And we assume that infants regularly follow the gaze of observed speakers—an ability which also develops rapidly during this period (Baldwin et al. 1996).

An example of the network's input data is shown in Table 1. We assume that the infant is able to identify simple episodes taking place in her environment, and to represent these as structures of concepts, perhaps using the kind of scheme illustrated in Fig. 1a, each involving an agent performing an action. As the infant observes the world, she evokes a sequence of conceptual representations: at t_1 she observes daddy laughing, at t_2 she observes a cookie, at t_3 she observes a dog jumping, and so on. At the same time, the infant is hearing words produced by mature speakers in the environment.

In this input data, we introduce a weak correlation between words and conceptual representations. We assume the data represent a succession of situations, each of which contains a subset of the set of agents, performing a subset of the set of possible actions. The words heard by the infant in a given situation are more likely than chance to refer to the agents and actions in this situation, because speakers regularly talk about objects and actions in the current situation—so through a process of cross-situational learning (Siskind 1996), the infant can slowly learn correct associations between concepts and words. But there is a large amount of noise in the mapping between concepts and words in the input data, so this cross-situational learning is very inefficient.

We also introduce another kind of regularity in the input data which provides a better opportunity to learn concept-word mappings. We assume that infants routinely follow the gaze of speakers, and that speakers often look at the objects and events they are talking about if they are physically present (see, e.g., Yu and Ballard 2007). If the infant happens to apprehend an episode in which a mature speaker is talking, there is a brief moment afterwards when the correlation between concepts and words in the infant's mind is much stronger than usual: The infant is more likely than usual to perceive an episode that the speaker is talking about. This is illustrated at t_4 and t_5. At t_4, the infant apprehends her mother talking (MUMMY TALK). The infant then follows the speaker's gaze, and at t_5, perceives an episode in which a cat runs (CAT RUN). At the same time, in the input medium representing incoming word forms, the infant is representing the words produced by the speaker, in particular the words *cat* and *run*.[1] The moment just after the infant perceives a speaker talking, therefore, constitutes a particularly good opportunity to learn a mapping between words and concepts. For the infant, recognition of a talk action can be thought of *as a cue to engage in some word-meaning learning*.

It is important to say a little more about what is going on at times t_4 and t_5. At t_4, the infant observes a physical action, which in some ways is very similar to other motor actions such as jumping or laughing: an agent is producing certain gestures which are of a recognizable type. In our model, the action type TALK simply represents a certain type of motor action: in itself it does not encode any of the special properties of communicative actions that were discussed in Sect. 2. However, the talking action is special in that it is represented twice by the infant: once as a semantic concept denoting a motor action taking place in the world, and once in a special medium holding motor actions that potentially convey meanings—namely, the medium which holds word forms. We assume that this medium automatically processes phonological signals picked up by the infant. At the moment, when the infant attends to a talking event (t_4), the signals encoded in this medium are constrained to be those produced by the speaker, because the semantic and phonological representations are derived from the same perceptual input. At the next moment (t_5), when the infant establishes joint attention with the speaker, we assume that the representations of the word forms produced by the speaker remain active, in some form of phonological working memory (see, e.g., Baddeley et al. 1998): Thus, in Table 1, the speaker's words *cat* and *run* are active at both t_4 and t_5.

[1] Our simulation considers only content words: We do not consider the issue of how the meanings of function words are learned or how the infant learns the syntactic principles that map surface sequences of words with episode representations. But these issues are the focus of a separate neural network model (see Takac et al. 2012).

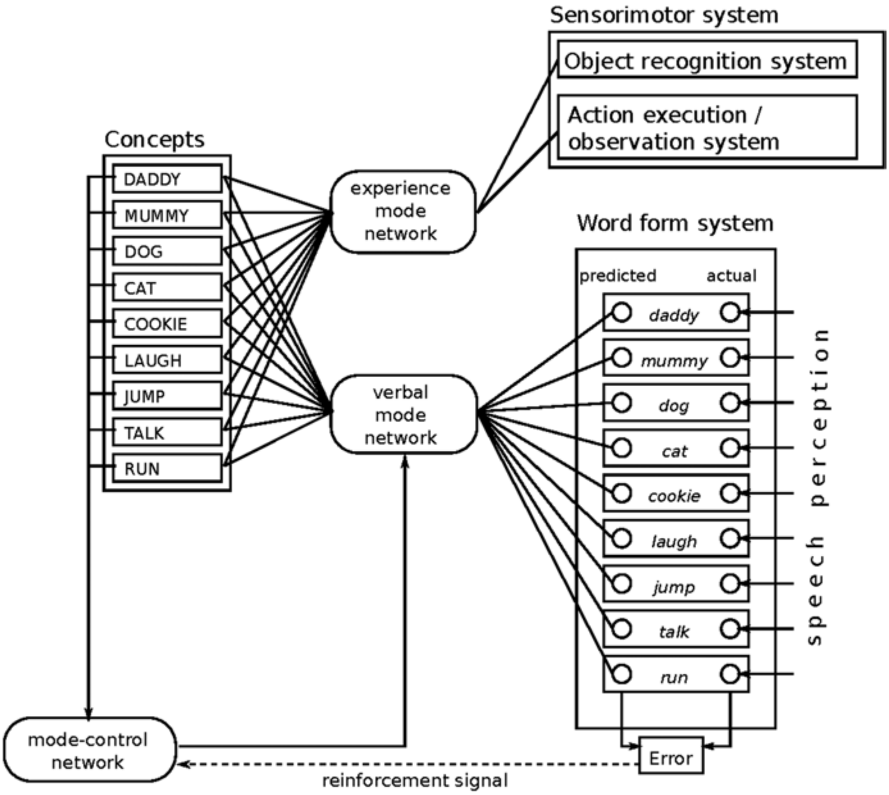

Fig. 2 Architecture of the word-learning network

3.2 Network Architecture

The network's architecture is shown in Fig. 2. In this section, I will describe the two key features of the architecture.

3.2.1 Experience Mode and Verbal Mode Networks

One key feature of the network architecture is that conceptual representations (DADDY, MUMMY, RUN, JUMP, etc.) are linked separately to the sensorimotor system and to word forms. The *experience mode network* links concepts to the agent's perceptual and motor interfaces with the world. During normal experience, this network is engaged: through this network, when the agent perceives objects in the world, or activates motor programs, this activates conceptual representations. However, we also envisage a separate network, the *verbal mode network*, which links the agent's conceptual representations to a specialized neural medium encoding word forms, or some other repertoire of atomic communicative gestures (for instance, hand gestures in sign language). Our proposal is that the infant can *selectively* engage "verbal mode," by turning on the connections in this network. When verbal mode is disengaged, the stream of word forms arriving in the phonological input buffer is effectively ignored by the infant. When verbal mode is engaged, the infant can learn associations between currently active concepts and word forms—and later, when associations have been learned, the infant can use word forms to activate concepts by themselves.

3.2.2 The Mode-Control Network

A second key feature of the architecture is a network which learns *when* to engage verbal mode, as a function of current experiences. We propose that there are certain moments when it makes particular sense for the infant to engage verbal mode. In particular, it makes sense to engage verbal mode *immediately after perceiving a talking action*. As discussed in Sect. 3.1, the perception of a talking action signals an imminent word-learning opportunity. After the infant observes a talking action, there is a brief period of time during which there is a particularly reliable mapping between active concepts and active word-form representations: For the infant, this is an ideal moment to do some word learning, and therefore an ideal moment to engage verbal mode.

In our model, the infant learns to engage verbal mode through *reinforcement*, in roughly the same way she learns when to execute ordinary motor actions (Sutton and Barto 1998). In regular operant learning, the agent experiences a sequence of perceptual stimuli, and is taught a specified mapping from these stimuli to motor responses by a *reward schedule* that rewards the agent whenever a particular action follows a particular perceptual stimulus. Initially, the agent executes actions from her motor repertoire at random. From time to time she executes an action which results in a reward: when this happens, she learns an association between this action and the perceptual stimulus that preceded it, so that the next time this stimulus appears, she is more likely to execute the associated action.

The task of learning when to engage verbal mode is performed by the *mode-control network*. When verbal mode is engaged, the concepts that are currently active are mapped to *predicted word forms*. These predictions are compared to the *actual word forms* active in the phonological input buffer, and an error term is generated reflecting the accuracy of the prediction. From this error term a reward signal is generated, which is used to train the mode-control network. If the error is low, the signal is a reward; if it is high, it is a punishment. (The magnitude of the punishment is relatively small compared to that of the reward.) The structure of the input data, together with this reward schedule, cause the mode-control network to learn to engage verbal mode immediately after perceiving a talking action, and in no other circumstance.

3.3 Learning in the Verbal-Mode and Mode-Control Networks

At the start of training, weights in the mode-control and verbal-mode networks are initialized to random values. During training, the output of the mode-control network is annealed with noise, so that it initially engages verbal mode at random, but over time comes to deliver an output based on the activity of the concept units which provide its input.

Learning proceeds as follows. To begin with, the mode-control network engages verbal mode at random, after perceiving arbitrary episodes (e.g., CAT JUMP). In almost all cases, this leads to a small punishment, because the verbal-mode network has not yet learned any correct associations between concepts and words. But there is enough randomness early in learning that the network is turned on quite frequently, nonetheless. When it is turned on, some cross-situational word learning occurs. This learning is very slow: The baseline mapping from concepts to words in the input data is extremely noisy, as discussed in Sect. 3.1; in addition, learning only happens at the moments when verbal mode is engaged, so the network is only exposed to a subset of the noisy input data. After a few correct concept-word mappings have been learned, however, there is a subtle change in the reward schedule. If verbal mode is engaged *after perception of a talk episode* (e.g., MUMMY TALK), there is an increased chance of a correct mapping between concepts and word forms at the next time point, and thus an increased chance of a large positive reward. The mode-control network thus starts to learn to engage verbal mode after perceiving talk episodes. Engaging verbal mode after perceiving other episode types continues to result in a small punishment, so the network begins to learn to engage verbal mode *only* after perceiving talk episodes. Once this happens, the verbal-mode network starts to receive more reliable training data, and it begins to learn words more efficiently. As it becomes better at predicting word meanings, talk episodes in turn become better predictors of reward, and the rate of learning increases in the mode-control network. In short, the verbal-mode and mode-control networks *bootstrap* one another.

Figure 3 charts learning in the network in two experiments. In the first experiment, learning happens in parallel in the mode-control and verbal-mode networks, as outlined above. In the second experiment, verbal mode is engaged at every time point, and the mode-control network is not used. The dotted line shows the percentage of words correctly learned at each epoch of training when the mode-control network is not used. Learning proceeds at a roughly constant rate in this condition. The dashed line shows the percentage of words correctly learned at each epoch when the verbal-mode and mode-control networks are bootstrapping one another. (The solid line charts learning in the mode-control network: the percentage of times this network engages verbal mode in response to a talk action.) In this condition, learning of words is initially slower because verbal mode is only engaged around half the time. But as the mode-control network learns to selectively enable verbal mode in response to perceived talk actions, the speed of word learning increases significantly. While word learning now only happens at a small proportion of time

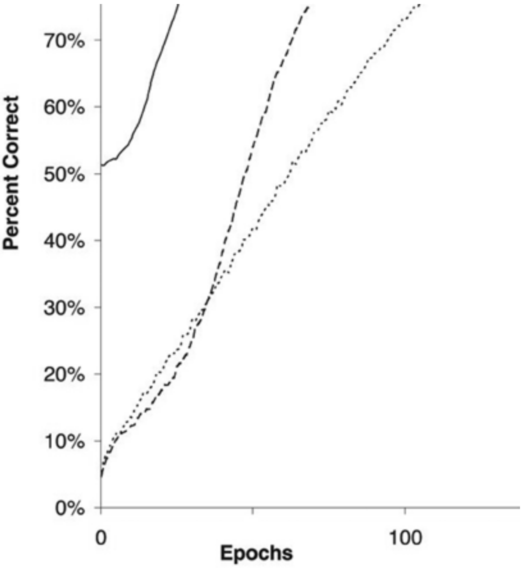

Fig. 3 Learning in the mode-control and verbal-mode networks in the model of Caza and Knott (2012). *Dotted line:* word learning performance when the mode-control network is not used. *Dashed line:* word learning performance when the verbal-mode and mode-control networks operate in parallel, and bootstrap one another. *Solid line:* learning in the mode-control network. (The *solid line* indicates the percentage of trials in which the mode-control network makes the right decision–i.e. selects verbal-mode after a talk action, and in no other circumstance)

points, learning is still more efficient, because it is focused on the time points which carry the best information about the mapping between concepts and words, namely, communicative actions. As I will discuss in Sect. 4, this transition serves as an interesting model of how the development of communicative action representations enables infants to begin learning words efficiently at around the age of 12 months.

3.4 Behavior of the Network After Learning

When training is complete, the mode-control network routinely engages verbal mode whenever a talk action is perceived, and the verbal-mode network reliably maps concepts onto the correct word forms. At this point, we envisage some changes in the way verbal mode works. During training, verbal mode is engaged simultaneously with experience mode, so that the concepts evoked by experience are also those associated with word forms. Once training is complete, we suggest that verbal mode and experience mode become *alternatives* to one another, so that if verbal mode is engaged, experience mode is disengaged. We also assume that the distinction between predicted and actual word forms disappears, and that the learned associations between concepts and word forms run in both directions, so that when verbal mode is engaged, word forms arriving in the phonological input buffer activate their associated concepts.

With these changes in place, there are two completely different ways of activating conceptual representations. In experience mode, concepts are activated by sensorimotor experience of the world, and incoming words are ignored. In verbal mode, concepts are activated by incoming words, through associations learned by the verbal-mode network during training, and sensorimotor inputs are ignored.

4 Representations of Communicative Actions in the Model

The neural network model outlined above offers an interesting account of the developmental processes taking place in infants at around the age of 12 months, which play a core role in Tomasello's theory of word learning. Tomasello's theory posits that infants only begin learning words in earnest when they acquire two social-pragmatic skills: an ability to establish joint attention, and an ability to identify and represent communicative actions and their underlying intentions. As discussed in Sect. 1, Tomasello does not consider in any detail what infants' representations of communicative actions look like, or the mechanism through which these representations help infants to learn word meanings. Caza and Knott's model makes some concrete proposals on both counts. In this section I will discuss these.

Firstly, what can we say about how communicative actions are represented in the neural network model? In one sense, clearly, communicative actions are simply represented as ordinary physical action episodes: MUMMY TALK is an episode, recognized through the same abilities to perceive objects and motor actions as are used to perceive ordinary episodes like DOG JUMP. But once the infant learns to routinely enter verbal mode after identifying a talk action, there is an interesting new element of structure to the pattern of concepts activated by a talk action. Each talk action activates a sequence of *two episode representations* in the conceptual system. The first representation is of the episode as a physical action (e.g., MUMMY TALK). This representation causes the mode-control network to engage verbal mode, so the next concepts to be activated will be those associated with the currently active word-form units. These units are of course encodings of the word forms making up the utterance that has just been represented—so the concept units that become active next will reflect the *propositional content* of this utterance. In brief, when learning is complete in the network, each utterance which is perceived will be represented in a sequence of two patterns of activation in the concept units: the first representing the utterance itself and the second representing its propositional content.

Recall from Sect. 2 that any neural network model of communicative action representations must solve two difficult problems. First, it must allow the representation of nested propositions (e.g., "Mummy said [the cat ran]"). Second, it must capture the elusive relation of intentionality between the outer proposition and the inner one (the fact that Mummy's saying action is "about" the cat running). The network of Caza and Knott, when trained, offers an interesting solution to both these problems. It represents a proposition nested within another proposition very simply, as a *sequence of two propositions*. In our model, the conceptual system can only represent one proposition (i.e., one episode) at a time: the propositional content of an utterance is represented entirely separately from the fact of the utterance itself, at the time point immediately after the utterance itself is represented. While the propositional content of the utterance is represented separately, the distinctive relation of "aboutness" that links the utterance and its propositional content is also captured by the network. This relation is not represented declaratively—rather, it is captured by constraints on how episode representations can succeed one another in the trained network. These constraints are partly due to the network's own internal mechanisms: when a talk action is perceived, the network is constrained to engage verbal mode. But they are also partly due to the way the network interfaces with the external world. The same perceptual stimuli which are perceived by the action recognition system as a talk action are encoded by the speech perception system as word forms—so when verbal mode is established after a talk action is perceived, the word forms that activate concepts are constrained to be those produced by the perceived speaker.

As discussed in Sect. 2, representations of propositional attitudes have two distinctive characteristics: I will now consider whether the representations of communicative actions in Caza and Knott's trained network have these characteristics. Firstly, the representer of a propositional attitude is committed to the *fact* of the attitude (e.g., the fact that a given agent has a given belief), but not to its propositional content. Do the network's representations of communicative actions have this property? Consider an example of communicative action "Mummy says [the cat jumps]." Note that the trained network evokes the episode representation MUMMY TALK directly from sensory experience, but this is not the case for the representation CAT JUMP: this representation has no relation to the network's sensory experience at all. The network does not implement any formal treatment of commitment, but if we assume a simple model, in which the network is only committed to the truth of the episodes it establishes in experience mode, then its representations of communicative actions correctly avoid commitment to the propositional content of communicative actions. Secondly, statements about propositional attitudes are intensional: their truth depends on the words that convey their propositional content. Communicative actions represented by the network certainly have this property. Their propositional content is activated in verbal mode, through associations between words and concepts. Even if a reliable observer knows that two words happen to designate the same individual in the world, there is no necessity that these words map onto the same concept in some arbitrary network: whether or not this is the case depends on the precise training that this network has received. Thus, the network's representations of communicative actions seem to have many of the right properties to qualify them as representations of propositional attitudes.

Secondly, Caza and Knott's model suggests a neural mechanism through which infants' pragmatic development supports their word-learning abilities. Recall that Tomasello proposes infants must acquire pragmatic skills before they can efficiently learn word meanings, but does not suggest *how* the development of pragmatic representations enables words to be more efficiently learned. Caza and Knott's model makes a specific proposal about how pragmatic learning about the special status of communicative actions impacts on the efficiency of word meaning learning.

To summarize: Caza and Knott's network extends Tomasello's social-pragmatic theory of infant word learning in two ways. Firstly, it provides the basis for a novel model of how communicative actions are represented, that goes some way towards capturing their distinctive properties as conveyors of propositional content.

Secondly, it provides an account of the mechanism via which infants' understanding of communicative actions supports their learning of word meanings.

5 Discussion

5.1 Communicative Action Representations as Instances of Semantic Representations

It is interesting to compare the model of communicative action just proposed to other accounts of cognitive representation. The model shares several features with other more general proposals about the form of cognitive representations. A particularly interesting point of contact is with Ballard et al.'s (1997) model of *deictic routines*. Ballard et al. argue that the cognitive representations active in an agent at any given moment in time cannot be interpreted in isolation, as they often "implicitly refer" to the agent's momentary deployment of perceptual and motor resources to his immediate environment. For instance, neural assemblies in the visual object categorization system in inferior temporal cortex predominantly represent the stimulus at the current fixation point, or at the current locus of covert attention (Zhang et al. 2011): in order to interpret these representations, we need to know what attentional action resulted in their activation. For Ballard et al., cognitive representations are often given meaning by their position in a sequentially structured routine of cognitive operations—for instance, a routine in which an agent attends to an object, then computes its grasp affordances, and then reaches for it. Each cognitive operation in the routine generates transitory cognitive representations—and often transitory motor states—which provide the conditions under which the next cognitive operation can be executed. The sequence of episode representations which collectively encode a communicative action in the current proposal can usefully be thought of as a deictic routine of this kind. The episode representation MUMMY TALK, when active, enables execution of a cognitive operation which changes the way the cognitive system is deployed to the world; it is impossible to interpret the conceptual representations which are activated next in the network without making reference to this cognitive operation, and to the episode representation that triggered it, even though this is no longer active.

Ballard et al.'s model of deictic routines is extended by Knott (2012), who proposes that an agent perceives *all* concrete episodes through sensorimotor routines with canonical sequential structure, and represents all such episodes as prepared sensorimotor routines. (A computational model is provided by Takac and Knott 2013.) In this account, semantic representations of concrete episodes are uniformly structured as sequences, and many principles of syntax are seen as deriving from constraints in the way sensorimotor operations can succeed one another sequentially. The idea that an utterance and its propositional content are represented at two distinct moments in time might seem unusual for theorists accustomed to thinking of semantic representations as static patterns of activity. But in the light of accounts like those of Ballard et al. (1997) and Knott (2012), communicative action representations are not exceptional in having sequential structure, but actually conform to the general pattern for semantic representations of episodes.

5.2 Towards a General Model of Propositional Attitude Representations

As noted in Sect. 1, infants do not develop sophisticated representations of propositional attitudes immediately: the early representations of communicative actions posited by Tomasello as having a role in word learning are presumably simple precursors to the representations that develop later in life. Indeed, the model of propositional content representations put forward in the current chapter is just a model of communicative actions: It is tailored to these actions and does not attempt to provide an account of other types of mental state, such as belief or desire. However, since it provides a novel account of the propositional content of utterances, it is interesting to consider whether it can provide the basis for a more general model of mental states. I will conclude with a few suggestions about this prospect.

Firstly, the idea that a propositional attitude representation in general takes the form of a sequence of two simple proposition representations, separated by some mode-changing operation, is an interesting one. In a more general model of propositional attitudes, the suggestion would be that the first proposition in the sequence would encode the agent adopting the attitude, together with a special action or operation denoting the attitude in question: BELIEVE, WANT, REMEMBER, and so on. In each case, activation of this special representation would trigger a change in cognitive mode, resulting in the activation of a new proposition in the conceptual system. Different attitudes would presumably trigger different cognitive modes: perhaps WANT would configure the conceptual system to represent an intention of the agent rather than the results of sensorimotor experience, for instance, while REMEMBER would configure the conceptual system to receive episode representations from long-term memory. I assume there would have to be hard-wired circuitry in the network to support each of the distinct propositional attitudes. In the case of REMEMBER, there is actually good evidence that memory retrieval involves the establishment of a special cognitive mode, implemented by specialized neural circuits (see, e.g., Buckner and Wheeler 2001; Buckner et al. 2008).

Secondly, a more complete model of attitudes would need to allow the agent to represent his own attitudes as well as those of other agents. In the account of communicative action representations given in this chapter, the communicative actions are always those of an observed agent. But the infant herself can execute communicative actions: how would the current model need to be extended to accommodate an account of producing utterances as well as of producing them? In an action execution scenario, we have to imagine that the infant uses representations in the conceptual system to plan her own actions—and that these actions can include the action of entering verbal mode for the purposes of speaking. Here, presumably, we must envisage a process which actively removes the infant's planned action of entering verbal mode (ME TALK) as soon as it is achieved, and replaces it with a representation of the content to be produced. Where this content comes from is an interesting question. We must also envisage that conceptual representations activate word forms in this mode, and that word forms result in overt speech sounds. In the case of a more abstract propositional attitude like wanting or remembering, it is likely that the operations evoking one's own attitudes are somewhat simpler than those evoking those of another agent: for instance, while evoking one's own desire in the concept units might just involve activating an interface to these units from one's own planning system, evoking the desire of another agent is likely to require perceptual inference mechanisms, and perhaps also specialized mechanisms for storing mental states of other agents.

A final interesting question concerns whether a mode-changing model of mental states supports arbitrarily deep nesting of mental states. For instance, imagine an agent who enters verbal mode (ME TALK) and then evokes the representation of another mode-changing operation—for instance ME WANT. Given the agent is in verbal mode, the operation WANT is presumably mapped to a word (*want*)—but is the WANT operation also executed in this mode?

All in all, the mode-changing model of communicative action representations seems to provide quite an interesting platform for the development of a more elaborate account of the representation of propositional mental attitudes. But it certainly raises more questions than it answers.

References

Averbeck, B., Chafee, M., Crowe, D., & Georgopoulos, A. (2002). Parallel processing of serial movements in prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(20), 13172–13177.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*(1), 158–173.

Baldwin, D., Markman, E., Bill, B., Desjardins, R., Irwin, J., & Tidball, G. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child Development, 67,* 3135–3153.

Ballard, D., Hayhoe, M., Pook, P., & Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*(4), 723–767.

Brentano, F. (1874). *Psychologie vom empirischen Standpunkte*. Leipzig: Duncker & Humblot.

Buckner, R., & Wheeler, M. (2001). The cognitive neuroscience of remembering. *Nature Reviews Neuroscience, 2,* 624–634.

Buckner, R., Andrews-Hanna, J., & Schacter, D. (2008). The brain's default network: Anatomy, function and relevance to disease. *Annals of the New York Academy of Sciences, 1124,* 1–38.

Butterworth, G., & Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms for perspective taking in infancy. *British Journal of Developmental Psychology, 9,* 55–72.

Caza, G., & Knott, A. (2012). Pragmatic bootstrapping: A neural network model of vocabulary acquisition. *Language Learning and Development, 8,* 1–23.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science, 26,* 609–651.

Dennett, D. (Ed.). (1989). *The intentional stance*. Cambridge: MIT Press/Bradford Books.

Diesendruck, G., Markson, L., Akhtar, N., & Reudor, A. (2004). Two-year-olds sensitivity to speakers' intent: An alternative account of Samuelson and Smith. *Developmental Science, 7*(1), 33–41.

Gärdenfors, P. (2004). *Conceptual spaces*. Cambridge: MIT Press.

Jacquette, D. (Ed.). (2004). *Brentano's concept of intentionality*. Cambridge: Cambridge University Press.

Knott, A. (2012). *Sensorimotor cognition and natural language syntax*. Cambridge: MIT Press.

Miller, E., & Cohen, J. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167–202.

Montague, R. (1974). The proper treatment of quantification in ordinary English. In R. Thomason (Ed.), *Formal philosophy: Selected papers of Richard Montague (pp. 247–270)*. New Haven: Yale University Press.

Plate, T. (2003). *Holographic reduced representations. CSLI Lecture Notes Number 150*. Stanford, CA: CSLI Publications.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928.

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61*(1-2), 39–91.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction. Cambridge*. Massachusetts: MIT Press.

Takac, M., & Knott, A. (2013). A neural network model of working memory for episodes. In M. Knauff (Ed.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 1432–1437). Berlin: Curran Associates, Inc.

Takac, M., Benuskova, L., & Knott, A. (2012). Mapping sensorimotor sequences to word sequences: A connectionist model of language acquisition and sentence generation. *Cognition, 125,* 288–308.

Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & P. Dunham (Eds.), *Joint attention: Its origins and role in development (pp. 103–130)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics, 10*(4), 401–413.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition. Cambridge*. MA: Harvard University Press.

Tomasello, M., & Herrmann, E. (2010). Ape and human cognition: What's the difference? *Current Directions in Psychological Science, 19*(1), 3–8.

van der Velde, F., & de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences, 29,* 37–108.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing, 70*(13-15), 2149–2165.

Zhang, Y., Meyers, E., Bichot, N., Serre, T., Poggio, T., & Desimone, R. (2011). Object decoding with attention in inferior temporal cortex. *Proceedings of the National Academy of Sciences of the USA*, 108(21), 8850–8855.

What Action Comprehension Tells Us About Meaning Interpretation

Tzu-Wei Hung

Abstract This chapter defends the view that the functional mechanisms underlying the capacity of understanding the means-end structure in action are necessary and in some cases sufficient for the capacity of understanding the symbol-referent relationship in language. This chapter first examines the relationship between action and speech and cognitive requirements that are needed by symbol-referent mapping but not means-end mapping. It further explores the mechanisms that are indispensable for means-end mapping and investigates whether they are sufficient to explain symbol-referent mapping at the most basic level. Finally, it describes how this analysis is consistent with data in both animal communication and in children with and without semantic pragmatic language disorder.

1 Introduction: Defining the Questions

This chapter defends the view that the functional mechanisms underlying the capacity of understanding the means-end structure in action are necessary and in some cases sufficient for the capacity of understanding the symbol-referent relationship in language. The motivation of this chapter is that our ability to comprehend and generate *instrumental actions,* i.e., intentional behaviors with means-end structures, is believed to be highly relevant to the ability to understand and produce language (Byrne 2006; Garrod and Pickering 2008; Kiverstein and Clark 2008; Wolpert et al. 2003). Some researchers suggest that language itself is a component of sensorimotor action (Henis and Levinson 1995), that sensorimotor circuits form a cortical basis for language (Pulvermüller and Fadiga 2010), or that a sensorimotor model can account for syntactic processing (Knott 2012). The extent to which the capacity for understanding action also facilitates the understanding of language is therefore an interesting question.

Hurley (2006, 2008) once noticed that a flexible and arbitrary relationship between means and ends in action comprehension provides a basis for a flexible and arbitrary relationship between symbols and referents. Unfortunately, Hurley did not provide further explanation before she passed away in 2007. A prima facie thought

T.-W. Hung (✉)
Institute of European and American Studies, Academia Sinica, Taipei, Taiwan
e-mail: htw@gate.sinica.edu.tw

Tzu-Wei Hung (ed.), *Communicative Action,* DOI 10.1007/978-981-4585-84-2_8,

is that understanding the two relationships involves similar cognitive skills and hence corresponding mechanisms. For example, to grasp the means-end structure of an action, an observer must first perceive something as an action and then recognize this action as *the means* to achieve a certain goal. The former requires a mechanism of segmenting constituents (movements or actions) from a continuous visual flow while the latter relies on a mechanism linking the action with the effect that the performer intends to bring about. Likewise, without the mechanism that segments the constituents (morphemes or words) from a continuous auditory flow and without the mechanism that associates a phonetic word with whatever the speaker may intend, a hearer cannot comprehend symbol-referent association. Hence, it seems that these mechanisms are necessary for means-end mapping as well as symbol-referent mapping.

However, explaining symbol-referent mapping may not be so simple because the relationship between the two mappings is unexamined, as is the question of whether the latter requires cognitive skills beyond those required by the former. Hence, the central proposal cannot be defended without clarifying the following questions:

1. What is the relationship between action and speech, and more specifically, what is the relationship between means-end structure and symbol-referent structure?
2. What mechanisms are required to link the means with an end?
3. Do these mechanisms enable symbol-referent association?

To narrow down the scope of the investigation, the answers should be sought mainly at the subpersonal-functional level in which symbol-referent mapping is clarified in terms of the causal and mechanical correlation among functionally individuated components. Other levels of explanation, such as computational or neural implementation, albeit important, are far beyond the scope of this work. To this end, Sect. 2 starts with the comparison between action and speech and between means-end and symbol-referent mapping. Section 3 examines the mechanisms that are indispensable for linking the means with ends of an action. Section 4 discusses whether the same mechanisms facilitate symbol-referent mapping. Finally, Sect. 5 describes how this analysis is consistent with data in both animal communication and in children with and without semantic pragmatic language disorder (SPLD)—the impairment of understanding semantic and pragmatic aspect of language.

2 The Relationship Between Action and Speech

The relationship between action and speech can be illustrated in terms of the relationship between various types of action (Fig. 1). Actions are primarily instrumental and noninstrumental. The former are intentional behaviors with means-end structures while the latter are those without the structures (e.g., bodily expression of emotion). Instrumental actions may involve collaboration (e.g., moving lumber with others) or they may not (e.g., crushing a nut with a hammer). Collaborated instrumental actions include communicative and noncommunicative actions. A

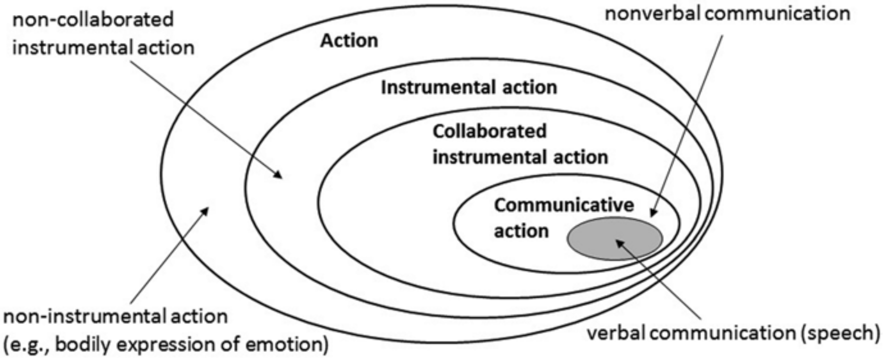

Fig. 1 The relationship among types of action

communicative action is the mixture of physical movements to achieve a certain communicative end. By definition, it is a special case of instrumental action. Communicative action can be decomposed into nonverbal (gesture) and verbal communication (speech). Therefore, speech belongs to action, which motivates one to apply the mechanisms of understanding the former to understand the latter.

More specifically, means-end association of action resembles symbol-referent association of language in several aspects. For example, they both exhibit *hierarchical relationship*. Phonetic segments (morpheme or word) can form higher-level units (phrase or clause) and even higher units (sentences). This is just as simple movements (grasping a hammer) can integrate into complex actions (nailing a lit on a box) and larger scales actions (making a bookcase). Besides, the two associations are *flexible*. While an action (waving hands) can serve as the means to achieve multiple goals (greeting or expelling mosquito), a different action can be used as the means to meet the same goal. This is similar to a word that may have different referents (ambiguity) and different words may be linked to the same referent (synonym). Moreover, similar cognitive skills are involved in understanding actions and words. For example, the skill of *intention detection* is vital for both. When associating an action with a goal, just as when associating a symbol with a referent, an observer needs to detect the intention of the actor/speaker and infer the goal from observable action/speech in everyday life. Understanding collaborated instrumental action and verbal communication involves both joint attention (a link-up between an observer's perception of actor's focus on a movement and the observer's focus on the same event) and ritualization (in which organisms in repeated instances of interaction shape communicative signals). In both cases, detecting other's intention is the key for tracking other's attention and mutual intention. Likewise, both associations require *inferential process* to understand a novel action/word or a familiar action/word in unfamiliar situation based on known clues. Accordingly, mean-end mapping and symbol-referent mapping share similar features and cognitive capacities. If the former requires a domain-general mechanism to enable these features and capacities, the latter should also require the same mechanism to enable the same features and capacities. Therefore, in this sense, the mechanism of means-end mapping is *necessary* for symbol-referent mapping.

Nonetheless, symbol-referent mapping also exhibits significant differences and involves additional cognitive skills at least in the following aspects. First, the object or event to which a symbol refers may not be easy to observe. To understand instrumental action, if a learner notices a salient event (a nut is being crushed) following an action (using a hammer), he/she may infer the tool's function simply based on its physical properties (hard enough to crush a nut) (Moore 2013). Conversely, there is no way for a hearer to infer the meaning of novel word in an unfamiliar language merely based on his/her perception of the sound of that word. Other clues, such as observation and previous knowledge of nonverbal interaction with the speaker or the conjecture of speaker's intention, are important. Another similar but different case is that a symbol may refer to rather abstract entity (god, love, and peace) that has no concrete referent at all. Without first recognizing the effect of a communicative action, an observer can hardly infer the means-end structure. In both cases, the inferential capacity of action comprehension helps little. The observer must also rely on a sophisticated detector of intention, which is a prerequisite for processing and learning of speech (Over and Gattis 2010). Third, the symbol-referent connection is more arbitrary and flexible than is means-end connection. On the one hand, it is more arbitrary because not all physical producible phone sequences make words and not all word sequences qualify as sentences. Symbol-referent connection should generally conform to linguistic conventions. On the other hand, it is more flexible because while the connection is basically confined by conventions, a speaker can also freely alter the association in somewhat unconventional way. Thus, a hearer should be able to differentiate these highly flexible and arbitrary ways of connection. Fourth, symbol-referent mapping requires word grouping. A competence hearer can know a novel word's referent through inferring, but this cannot be possible without first sorting word into different categories. Word grouping is useful not only for learning words effectively, but also for reapplying learned rules of combination (e.g., morphological or syntactic rules) to words.

Accordingly, the mechanism of means-end mapping should also facilitate the symbol-referent mapping of above cognitive skills; namely, particularly in terms of *advanced intention detection, arbitrary and flexible mapping,* and *word grouping*. In this sense, a mechanism that can enable these skills is *sufficient* for symbol-referent mapping at the most basic level.

3 The Mechanism of Means-End Mapping

To identify the goal of an action, the observer must causally associate the actor's movements with the effect of a change that the actor intends to bring about. The intention is for the actor's mental state to drive the performance of an action to achieve a goal, whereas the goal is the physical change in the world that the actor desires. Knowing one helps reveal the other. However, the manner in which the association between action and goal is established at both the functional and neural implementation levels is the subject of debate.[1] Here, we presuppose Brass et al.'s (2007) analysis to explain the action goal, assuming that understanding the goal in an unfamiliar situation depends more on the observer's inferential processes, whereas simulative mirroring is more important in a familiar situation. Following this line of thought, if the mechanism recognizes a limited range of suspected goals in a newly observed action, it can then select fewer goals from among the possibilities.

Then, what functional mechanism is required to establish the link? When performing an action, an actor's cognitive system must activate multiple motor controllers to induce a series of bodily movements to complete the entire action. However, selecting controllers under given bodily states and environmental context is problematic. To solve this problem, Wolpert et al. (2003) proposed a motor selection mechanism called the modular selection and identification for control (MOSAIC). The MOSAIC contains multiple controller-predictor pairs as basic elements (Fig. 2a), with each controller generating not only a motor command to activate corresponding bodily movement, but also an efference copy of motor command. The efference copy is sent to the paired predictor to simulate the outcome of that command so that the simulative prediction can be sent to comparators (represented by AND "⊗" and OR "⊕"gates) for signal calibration. The MOSAIC runs numerous controller-predictor pairs concurrently. If the comparators indicate that the difference between a predictor's simulative prediction and a controller's efference copy is within a given error range, the controller will be chosen. Otherwise, the controller will be dismissed. The comparison results are then stored in a memory component in terms of probabilistic rules. Consequently, motor command that is suitable under given environment can be chosen.

The MOSAIC models can also work together to achieve more accurate selection by allowing bidirectional and hierarchical motor control (known as the hierarchical MOSAIC (HMOSAIC), Fig. 2b). According to Wolpert et al. (2003), despite differences in the effectors and dynamics among various types of pens, humans can produce a number of compensating movements to preserve the kinematics of writing across different instruments. This situation suggests the existence of high-level reference signals (e.g., intentions) that have many ways of activating low-level controllers. When using Bayesian terms to describe cross-level communication in the HMOSAIC, controllers at higher levels receive posterior probabilities

[1] The<Emphasis Type=cQuotecItaliccQuotec> *direct matching hypothesis*</Emphasis> states that mirror neurons enable the simulation and recognition of the means-end structure by transforming the observer's sensory stimuli of observed actions into motor commands to execute that action (Gallese and Goldman 1998; Rizzolatti and Craighero 2004; Rizzolatti and Sinigaglia 2010), whereas the <Emphasis Type=cQuotecItaliccQuotec>*inferential reasoning account*</Emphasis> claims that the action goal is grasped through inferential interpretive processes with minimal help from mirror neurons (Csibra 2007; Gergely and Csibra 2003; Uithol et al. 2011). This dispute likely occurs because the concept of action contains multiple levels of complexity, resulting in the observed diversity of empirical findings.

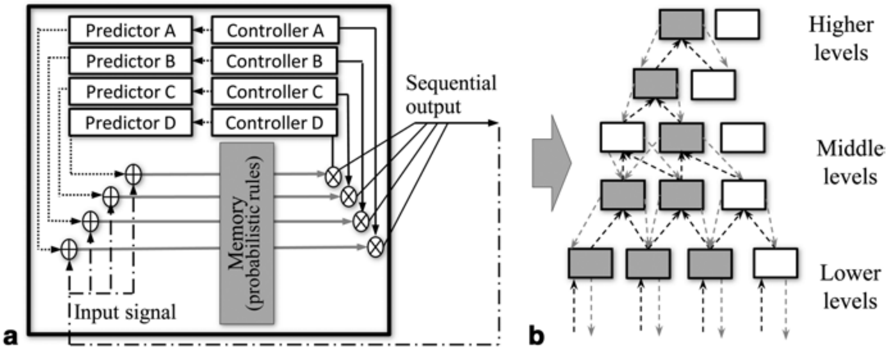

Fig. 2 **a** The MOSAIC model. **b** The HMOSAIC: Each *box* represents a MOSAIC. The *grey boxes* represent selected MOSAIC

from MOSAICs only at subordinate levels. Predictors at higher levels generate prior probabilities for MOSAICs at subordinate levels.

Although the HMOSAIC has originally been proposed for motor selection, it is also helpful for identifying the action goal. Technical details of identifying the goal may be laborious, but the principle is simple because the mind can establish the association by predicting, comparing, and revising.

To identify the goal in an unfamiliar situation, the HMOSAIC's higher level may randomly generate a controller-predictor pair to produce predictions about an actor's intended goal. These predictions are tested downward with input signals. When the comparator ⊕ indicates a gap, the predictions are revised to minimize the difference. This inferential processing helps narrow down the number of possible goal predictions. Likewise, the HMOSAIC allows bottom-up processing for direct matching. In unfamiliar situations, direct matching alone is insufficient to identify an action goal (Carpendale and Lewis 2008; Preston 2008; Uithol et al. 2011). However, when higher-level predictions that are activated by direct matching are cross-compared with those from inferential processing, the mechanism detects the actor's intended goal in a more precise way. For instance, suppose that some familiar segments of a novel action activate an observer's controller-predictor pairs at multiple levels. The higher-level predictions are cross-compared with the results from top-down processing such that the predictions with the highest likelihood (i.e., predictions with minimal gaps) will be singled out from numerous initial predictions. Accordingly, identifying actor's intentions requires repetitive processing of predictions, comparisons, and revisions.

4 The Mechanism for Symbol-Referent Mapping

We have seen that Wolpert et al.'s (2003) HMOSAIC can be used to identify the most probable predictive goal of the actor. However, can the above mechanism be extended to symbol-referent mapping?

At the first approximation, suppose that the cognitive system also needs to reiterate the process of predicting, comparing, and revising to map a word correctly. The HMOSAIC may initiate several controller-predictor pairs for generating predictions, which may be biased and based on previous observation and knowledge. For example, the mechanism might accidently detect that a sound, say, "bye-bye" is followed by some statistically less variant effects (e.g., someone's leaving), helping the mechanism predictively map the sound onto certain effects and not others. In addition, these predictions benefit from previous knowledge of nonverbal communication.[2] This knowledge improves initial higher-level predictions for top-down processing, which, if not enough to reveal the speaker's intention, help narrow down the number of possible predictions about the referent.

Meanwhile, nonverbal actions accompanying the speaker's sounds (gestures and facial expressions) may trigger lower-level controllers to facilitate bottom-up processing. In some circumstances, nonverbal actions alone are enough to allow the listener to interpret the speaker's intention (e.g., foreign tourists who do not understand a single word that a local robber says can still comprehend that they are being mugged). The results of bottom-up processing can be cross-compared with the results of top-down processing to determine the most likely predictions. Thus, the referent of "bye-bye" is identified and the word is learned.

However, at least two problems arise from the above description. First, when a fluent speaker intentionally utters a sentence, not all words he/she uses are consciously selected and explicitly intended. Then, how can the mechanism spot the "intended" referent? The answer is that the hearer can still use the mechanism to understand the speaker's words because the words are linked to the words that the speaker would say if he/she were aware of word selection. Given that tracking others' intentions is important for joint attention and ritualization in both nonverbal and verbal communication, the intention detection mechanism is powerful. Digressively, this advantageous mechanism may extend from understanding human interaction to interactions between human and nonhumans (e.g., animals, the sun, and the wind). We suspect that this mechanism partly explains the general trend of anthropomorphism (i.e., personification) in children's learning and in novelists' depictions of the physical world.

The second problem concerns the way in which this seemly simple description can capture the highly flexible and arbitrary nature of word mapping, given that a word may have more than one connecting target (ambiguity) and different words

[2] In children's language acquisition, 12-month-old infants who are pointing out objects to adults exhibit a joint attention skill that is beneficial for language development (Tomasello et al. 2007). Eighteen-month-old infants rely on their shared experience with adults to interpret the meaning of adults' gestures during nonlinguistic communication (Liebal et al. 2009).

may share the same target (synonyms). Besides, while some words' connecting targets are statistically invariant across contexts (literal meaning), others vary among speakers and situations (contextual meaning). Thus, how does the mechanism differentiate various meanings, such as word meanings from speaker meanings? At the least, the four following types of meaning should be investigated:

a. Word meaning (lexical meaning of *word-type*), which is conventional and context-independent, e.g., "she" is the third person singular feminine pronoun, and "penguin" is an aquatic species living in the southern hemisphere.
b. Word meaning (lexical meaning of *word-token*), which is conventional but context-dependent, e.g., "she" may mean someone's mother or daughter, and "penguin" may refer to a species or an individual organism.
c. Speaker meaning (*conventionally* implicated), which is what a speaker intends to convey beyond his conventional use of words, e.g., replying, "I am married" to the query, "Can I have your number?", or answering, "I am Brazilian" to the question, "Do you play soccer?"
d. Speaker meaning (*unconventionally* implicated), which is what a speaker intends to convey beyond his unconventional use of words, e.g., a cleaner says to colleagues, "Check out the massive chocolate in the toilet," or someone who named his boat "Penguin" says, "My Penguin is sick."

To show how the mechanism learns the four types of meaning, let us first consider how an action can be learned through grasping its mean-end structure. An observer may see someone hitting an apple tree with a stick to get apples or hitting a lemon tree to obtain lemons. The observer may infer that hitting trees with a stick helps acquire what actors want (apples or lemons) and apply the action in various situations. The action's means-end structure and the actor's intention are intimate. When seeing someone approaching a tree with a stick, the observer may infer that the actor intends to obtain some fruit. However, the connection is not always so direct because an actor might hit a tree to hunt prey or pretend to be doing so just for fun or for deception. In these cases, the actor's intention cannot be derived from the action's structure alone.

Similarly, an observer may hear someone uttering "this" to indicate an apple on one occasion and "this" for a lemon on another occasion. The observer may infer that "this" maps onto the referent in (b) sense and may fine-tune the mapping across various situations to extrapolate (a) from (b). The word's mapping and the speaker's intention are intimate too. When hearing "penguin" and knowing what it conventionally maps to, the observer may infer that speaker intends to talk about a penguin. However, the connection is not always so direct, as a speaker might utter a word unconventionally, as in (d), or may utter the word conventionally but intend something else, as in (c). In these cases, the speaker's intentions cannot be derived from the word's mapping alone.

More specifically, through the HMOSAIC, a word has been tested and mapped to the referent that is more likely to match the linguistic input across a number of occasions. This predictive mapping is consistent with the conventional use of the word and is the correct prediction of word meaning in (a) sense. The mapping will be stored in memory and taken as a *default prediction* whenever the mechanism receives the same word in the future. If default predictions match both input words (e.g., "I am Brazilian") and other contextual clues (e.g., the speaker who says this is actually Brazilian), the HMOSAIC made a correct prediction of (b). Nonetheless, default predictions can easily fail. If a default prediction contradicts an input word, as in (d), then the mechanism needs the intention detection mechanism to revise its predictions. This adjusted prediction, when matched with contextual clues, results in the correct prediction of (d). Alternatively, if a default prediction matches an input word but seems irrelevant with regard to the contextual clues, as in (c), then the mechanism needs the intention detection mechanism again to modify its predictions, which, if matched, lead to the correct prediction of (c). In other words, the four meanings (a)–(d) can be differentiated and learned through comparing predictions, verbal input, and contextual input.

Furthermore, when a word is in its repertoire, the mechanism can use the stored mapping as a default prediction and test this prediction to identify the speaker's meaning. However, if an input word is novel and the word mapping has not been established, the mechanism needs to recognize the speaker's intention through nonverbal clues first and then map the word according to this intention. This explains why people sometimes, especially in language acquisition, infer the speaker's words (i.e., semantics) based on grasping the speaker's intention (i.e., pragmatics) and infer a communicator's goal (i.e., pragmatics) based on his/her words (i.e., semantics) in everyday conversation.

Grouping is another key feature of learning words. Without grouping, the hearer's mind can neither store words efficiently nor derive unknown words from known ones. During vocabulary expansion, the mind may detect some phonetically repeated patterns (e.g., fragmental sequences) of words, which reveal different similarities in different contexts and provide clues for sorting words. To determine whether words share similarities, the mind must compare words according to the same criteria. These criteria amount to reference signals of higher level MOSIAC and may depend on the observer's attention and the salience of the input properties. One way to compare words is to focus on their referents (i.e., semantic categorization) in which words referring to objects (i.e., nouns), motions (i.e., verbs), states of objects (i.e., adjectives), and states of motion (i.e., adverbs) can be differentiated. Another way is to focus on the ways in which repeated parts of phonetic sequences are segmented from entire sequences (i.e., morphological categorization). Thus, words with bound morphemes affixed to the head of word sequences (i.e., prefixes) will be separated from words with bound morphemes affixed to the end (i.e., suffixes). These two ways of sorting words are interrelated (e.g., words ending with -tion are frequently nouns), and they facilitate the hearer's generalization of a known word class to an unknown word (e.g., generating predictions in which any unknown word with -tion is mapped onto objects instead of motions). Finally, words can be sorted according to their position and function in a sentence (known as syntactic categorization). For example, the hearer's mechanism may detect that some words never sequentially follow certain others and that some bind two word sequences (e.g., conjunctions and prepositions). The MOSAIC's comparator will indicate whether words share similar positions and functions, and, if so, it will group them together.

To sum up, the mechanism for means-end mapping explains additional cognitive skills required for symbol-referent mapping (i.e., advanced intention detection, highly flexible and arbitrary mapping, and word grouping). This mechanism is thus the sufficient mechanism for symbol-referent mapping at least in the above-mentioned cases.

5 Empirical Supports

The view that the mechanism for means-end mapping is necessary and in some cases sufficient for symbol-referent mapping is consistent with previous reports. On the one hand, if the mechanism of means-end mapping is necessary for linking symbols with referents, then people who have the mechanism impaired are unlikely to have the capacity of symbol-referent mapping intact. The data from the studies of language disorder support this view. While SPLD is a class heterogeneous syndrome resulting from various causes, the deficit of meaning interpretation occurs frequently with the deficit of instrumental action comprehension. For example, it is reported that patients with Rett syndrome exhibit no means-end behavior beyond automatic responses to particular stimuli (Woodyatt and Ozanne 1992), and they have difficulty using word for a functional communicative purpose (Cass et al. 2003). Cass et al. (2003) found that only 18 of 84 of their subjects with Rett syndrome reported using words and only six of them used words in meaningful ways. Likewise, a high proportion of children with autism do not develop the skill to form and manipulate symbolic material (Prior and Ozonoff 2007), and the acquisition of the skill is argued to be associated with the skill of means-end reasoning (Abrahamsen and Mitchell 1990). On the other hand, if the capacity of means-end association is prerequisite for the ability to form symbol-referent association, then the former is unlikely to develop after the latter. This view is consistent with the fact that while infants can solve simple means-end problems, such as pulling a cloth to retrieve a toy, at as early as 6 months (Willatts 1999), they cannot look at the right portrait when hearing "Mommy" or "Daddy" until 6 months (Tincoff and Jusczyk 2012) and can only map meaning to newly segmented words at 17 months (Estes et al. 2006).

On the other hand, if the mechanism of means-end association is sufficient for linking symbols with referents, then organisms, which evolve with the former, should be able to complete the latter. However, this view does not imply that any animal capable of means-end mapping can be trained to understand human language. Because instrumental actions, as well as verbal communications, involve different level of complexity and hierarchy, organisms merely capable of comprehending lower-level means-end mapping (use a tool to get food) are less likely to possess the capacity of higher-level symbol-referent mapping (appreciate the unconventional implicated meaning of complex sentences). Rather, this view suggests that creatures capable of instrumental comprehension are also capable of some sort of communication. For example, capuchin monkeys (*Cebus apella*) can associate an action with a goal in problem solving (Yocom and Boysen 2010), and they could use token as a symbol too (Addessi et al. 2007). Asian elephants (*Elephas maximus*) are capable of means-end behavior (Irie-Sugimoto et al. 2008) as well as of recognizing the meaning of human trainer's sounds (Linden 2002). Millikan (2009) also argued that the subtle change in bee dance maps the nectar's location and serves as a descriptive sign that is isomorphic to the state of affairs. So based on this behavior, other bees can identify the direction, distance, and angle of the nectar. The dance maps the spatial details concerning the nectar for the observing bees and serves as a directive sign that is also isomorphic to the state of affairs. Accordingly, symbol-referent mapping is not unique to human species. Any species under the pressure of social interaction are likely to develop the ability to understand other's behaviors as well as signals (or natural signs).

6 Conclusions

To summarize, we first analyzed the relationships between action and speech and between means-end and symbol-referent mapping. We then examined the mechanisms that are indispensable for means-end mapping and discussed whether the same mechanisms also enable symbol-referent mapping. Finally, we discussed the consistency between our findings with the results reported for children both with and without SPLD as well as the results of studies in animal communication.

Symbol-referent mapping at higher level (complex sentences or paragraphs), in which a word's mapping may be adjusted according to the mappings of words in nearby sequences, may be more difficult to deal with. In this complicated case, whether the mechanism still sufficiently explains the symbol-referent mapping is another question that constitutes themes for future studies.

References

Abrahamsen, E. P. and Mitchell, J. R. (1990). Communication and sensorimotor functioning in children with autism. *Journal of Autism and Developmental Disorders, 20* (1), 75–85.

Addessi, E., Crescimbene, L., & Visalberghi, E. (2007). Do capuchin monkeys (Cebus apella) use tokens as symbols? *Proceedings of the Royal Society of London B*, *274*, 2709–2715.

Brass, M., Schmitt, R. M., Spengler, S., & Gergely, G. (2007). Investigating action understanding: Inferential processes versus action simulation. *Current Biology, 17,* 2117–2121.

Byrne, R. W. (2006). Parsing behaviour. A mundane origin for an extraordinary ability? In N. Enfield & S. Levinson (Eds.), *The roots of human sociality* (pp. 478–505). New York: Berg.

Carpendale, J. I. M., & Lewis, C. (2008). Mirroring cannot account for understanding action. *Behavioral and Brain Sciences, 31,* 23–24.

Cass, H., Reilly, S., Owen, L., Wisbeach, A., Weekes, L., Slonims, V., & Charman, T. (2003). Findings from a multidisciplinary clinical case series of females with Rett syndrome. *Developmental Medicine and Child Neurology, 45*(5), 325–337.

Csibra, G. (2007). Action mirroring and action interpretation: An alternative account. In P. Haggard, Y. Rosetti, & M. Kawato, (Eds.), *Sensorimotor foundations of higher cognition: Vol. 12. Attention and performance* (pp. 435–459). Oxford: Oxford University Press.

Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2006). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science, 18*(3), 254–260.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2,* 493–501.

Garrod, S., & Pickering, M. J. (2008). Shared circuits in language and communication. *Behavioural and Brain Sciences, 31,* 26–27.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences, 7,* 287–292

Henis, E. A., & Levinson, S. E. (1995). Language as part of sensorimotor behavior. *Proc. AAAI Symposium* (AAAI Technical Report FS-95–05). Cambridge, MA: Nov.

Hurley, S. (2006). Active perception and perceiving action. In T. Gendler & J. Hawthorne (Eds.), *Perceptual experience* (pp. 205–259). Oxford: Oxford University Press.

Hurley, S. (2008). The shared circuits model (SCM): How control, mirroring, and simulation can enable imitation, deliberation, and mind reading. *Behavioral and Brain Sciences, 31,* 1–22.

Irie-Sugimoto, N., Kobayashi, T., Sato, T., & Hasegawa, T. (2008). Evidence of means-end behavior in Asian elephants (Elephas maximus). *Animal Cognition, 11*(2), 359–365.

Kiverstein, J., & Clark, A. (2008). Bootstrapping the mind. *Behavioural and Brain Sciences, 31,* 41–52.

Knott, A. (2012). *Sensorimotor cognition and natural language syntax*. Cambridge: MIT Press.

Liebal, K., Behne, T., Carpenter, M., & Tomasello, M. (2009). Infants use shared experience to interpret pointing gestures. *Developmental Science, 12,* 264–271.

Linden, E. (2002). *The octopus and the orangutan: More tales of animal intrigue, intelligence and ingenuity*. New York: Plume.

Millikan, R. G. (2009). Biosemantics. In B. P. McLaughlin & A. Beckerman (Eds.), *The Oxford handbook of philosophy of mind*. Oxford: Oxford University Press.

Moore, R. (2013). Imitation and conventional communication. *Biology & Philosophy, 28*(3), 481–500. doi:10.1007/s10539–012-9349–8.

Over, H., & Gattis, M. (2010). Verbal imitation is based on intention understanding. *Cognitive Development, 25,* 46–55.

Preston, S. D. (2008). Putting the subjective back into intersubjectivity: The importance of person specific, distributed, neural representations in perception-action mechanisms. *Behavioural and Brain Sciences, 31,* 36–37.

Prior, M., & Ozonoff, S. (2007). Psychological factors in autism. In F. R. Volkmar (Ed.), *Autism and pervasive developmental disorders* (pp. 69–128). Cambridge: Cambridge University Press.

Pulvermüller, F., & Fadiga, L. (2010). Active perception: Sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience, 11,* 351–360.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27,* 169–192.

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nature Reviews Neuroscience, 11,* 264–274.

Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy, 17,* 432–444. doi:10.1111/j.1532–7078.2011.00084.x.

Tomasello, M., Carpenter, M., & Liszkowski, U. (2007). A new look at infant pointing. *Child Development, 78*(3), 705–722.

Uithol, S., van Rooij, I., Bekkering, H., & Haselager, P. (2011). What do mirror neurons mirror? *Philosophical Psychology, 24*(5), 1–17.

Willatts, P. (1999). Development of means-end behavior in young infants: Pulling a support to retrieve a distant object. *Developmental Psychology, 35*(3), 651–667.

Wolpert, D., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London B, 358*(1431), 593–602.

Woodyatt, G., & Ozanne, A. (1992). Communication abilities and Rett syndrome. *Journal of Autism and Developmental Disorders, 22*(2), 155–173.

Yocom, A. M., & Boysen, S. T. (2010). Capuchins (Cebus apella) can solve a means-end problem. *Journal of Comparative Psychology, 124*(3), 271–277.

Erratum

Chapter 4

When Actions Feel Alien—an Explanatory Model

Timothy Lane

Tzu-Wei Hung (ed.), *Communicative Action,* DOI 10.1007/978-981-4585-84-2_4,
© Springer Science+Business Media Singapore 2014

DOI 10.1007/978-981-4585-84-2_4

Timothy lane

Taipei Medical University, Institute of Humanities in Medicine and Shuang Ho Hospital, Brain and Consciousness Research Center

Academia Sinica, Institute of European and American Studies

National Chengchi University, Research Center for Mind, Brain, and Learning

Tzu-Wei Hung (ed.), *Communicative Action,* DOI 10.1007/978-981-4585-84-2_9,
© Springer Science+Business Media Singapore 2014