The Public Dimension of Meaning

Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Cathal Ó Madagáin
Department of Philosophy
University of Toronto

© Cathal Ó Madagáin 2011
Abstract: “The Public Dimension of Meaning”; Cathal Ó Madagáin, Doctor of Philosophy; Department of Philosophy, University of Toronto, 2011

The philosophical discussion of conceptual content and linguistic meaning in the 20th century has been dominated by two contrasting approaches - the descriptive-internalist approach, and the causal-externalist approach. Recent semantic models, for example the two-dimensional semantics of Jackson and Chalmers, attempt to integrate these two approaches. In this dissertation I explore a series of puzzles that highlight points at which the resources of these two approaches combined fall short. Particularly, the dissertation is an argument for the claim that facts about a linguistic community can affect the conceptual and linguistic content of individual members of that community, developing insights of theorists such as Quine, Wittgenstein, Kripke, Lewis and Davidson.

The study proceeds along two lines simultaneously, as an investigation into puzzles concerning conceptual content on the one hand, and concerning linguistic meaning on the other. The centerpiece of the investigation into linguistic meaning is a proposal for an irreducibly social aspect of linguistic meaning, which I call the ‘public content’ of linguistic terms. This proposal is motivated by the identification of some points at which neither individualist models of linguistic meaning nor the ‘social’ models of meaning currently available give convincing accounts. Drawing on recent developments in social epistemology, I argue that this aspect of meaning is determined by what speakers engaged in discourse would agree on under an ideal process of collective reasoning as the meaning of the terms they use. In the last chapter I attempt to reconcile this model of meaning with the two-dimensional semantic model, arguing for a three-dimensional model of meaning that includes internal, external, and public dimensions.

Alongside the discussion of linguistic meaning I explore a series of related puzzles that arise for conceptual content, particularly a new puzzle of referential indeterminacy, and the problem of conceptual error or normativity. I propose and defend solutions to these puzzles that lean heavily on the rational resources of individuals, focusing on the ‘personal level’ contents of thought to resolve puzzles in this domain, and rejecting models that lean on ‘sub-personal’ states such as neuronal, historical, or dispositional states of thinkers.
Table of Contents

TABLE OF CONTENTS III

THANKS IV

INTRODUCTION V

1: REFERENCE IN THOUGHT 1
1. CONTENT, REASON, AND MODALITY 1
2. PROBLEMS FOR DESCRIPTIVISM AND THE CAUSAL ALTERNATIVE 7
3. REFERENCE-FIXING DESCRIPTIONS 12
4. TWO-DIMENSIONALISM AND COMPLEX INTENSIONS 16

2: INITIAL BAPTISM AND THE QUA PROBLEM 21
1. THE QUA PUZZLE AND THE CAUSAL THEORY 21
2. THE QUA PUZZLE AND THE DESCRIPTIVE THEORY 23
3. WHY WE STILL HAVEN’T SOLVED THE PUZZLE 31
4. SHARED BELIEFS AND THE QUA PUZZLE 39

3: LINGUISTIC NORMS – FROM HUMPTY DUMMY TO HUNKY DORY 43
1. INTENTIONALISM AND ITS WOES 43
2. CONVENTIONALISM AND ITS WOES 50
3. NEW DIRECTIONS 55

4: CONCEPTUAL NORMS – THARTHRITE, QUUS, FLEEBEES 62
1. VARIETIES OF ERROR 62
2. SELECTING AMONG DISPOSITIONS 68
3. CATEGORICAL VERSUS HYPOTHETICAL IDENTIFICATION 76
4. LANGUAGE AND PUBLIC VERSUS PRIVATE CONCEPTS 84
APPENDIX: PHENOMENAL CONCEPTS AND COMPOSITION 90

5: PUBLIC CONCEPTS 95
1. VARIETIES OF DEFERENCE 95
2. PUBLIC KNOWLEDGE 105
3. PUBLIC CONCEPTS 111

6: THREE-DIMENSIONAL SEMANTICS 121
1. REFERENTIAL TERMS AND CONCEPTUAL CONTENT 121
2. SHARED BELIEFS AND LINGUISTIC CONTENT 123
3. THREE DIMENSIONAL SEMANTICS 125

CONCLUSION 130

BIBLIOGRAPHY 132
Thanks

Many thanks to everyone who helped me with this project. I have learned a huge amount from my supervisor Sonia Sedivy, and my committee members Bill Seager and Diana Raffman. Other philosophers at U of T and elsewhere have had a big impact on my work in the form of advice and encouragement, particularly: Benj Hellie, Suzanna Schellenberg, Brian Cantwell Smith, Peter Ludlow, Mohan Matthen, Stephen Biggs, David Chalmers, Tim Crane, John Searle, Angela Mendelovici, Peter Hartman and David Egan. Thanks to the University of Toronto for a five-year fellowship that allowed me to do this work, and to my family and friends for keeping me sane and happy over that time.
Introduction

This project establishes a model of linguistic meaning and conceptual content that pays particular attention to the role of the social in the determination of that content. Specifically, I argue that facts about at least pairs of speakers must be taken into consideration in order to understand an aspect of linguistic meaning that I call the ‘communicative value’ of linguistic terms. I ultimately argue that the meaning of terms in shared languages is not reducible to facts about any individual speaker of those languages, but can only be modeled on the basis of facts about the whole community. Along the way I explore various puzzles in the philosophy of perception and thought that arise as a result of my investigation into the roots of linguistic reference.

In the first chapter I present a model of referential terms and concepts (henceforth ‘concepts’) that can handle the twin challenges faced by twentieth century semantics – that of accounting for their psychological role, and also explaining the logical role of these terms in modal thoughts. I defend a version of what is known as two-dimensional semantics, where both internal facts about speakers’ minds and facts about their mind-independent environment must be taken into account to understand the referential content of their concepts. Over the course of the subsequent four chapters I identify points at which this model, which takes facts about individuals to exhaustively determine the content of their linguistic utterances, falls short.

The second chapter explores a puzzle that arises at the very foundation of linguistic practices – a puzzle for the introduction of new terms to a language. We can see, for example, a baby before us that we want to baptize. We point at the baby, perhaps, and say the name that we want to confer on the baby. Now what fixes the name that we baptize the baby with as a name for the whole person, and not just a time slice of the person, or the baby’s head? This puzzle is typically understood as a puzzle about the thoughts of the individual doing the baptizing, but I argue that a complete answer to the puzzle has to appeal to facts about not just the baptizer, but of those who witness the baptism, and subsequent users of the name. Here I introduce the idea of an aspect of
linguistic meaning that is essentially social, which I call the communicative value of language.

In chapter three I explore the two dominant models of linguistic meaning in the literature, and I present what I consider to be insurmountable difficulties for these theories. Particularly I focus on the phenomenon of linguistic normativity – the fact that we can use words with the wrong meanings, or in the wrong way. I revisit and expand on the notion of communicative value that I introduced in chapter two, and argue that it can provide us with an account of linguistic meaning that can handle the difficulties that standard models cannot, particularly by offering us an account of linguistic normativity.

In the fourth chapter I turn from linguistic norms to conceptual norms, and our apparent ability to apply our concepts in error. This phenomenon appears to play havoc with individualist theories of conceptual content – that is, most theories of conceptual content – but I argue that there is a way of thinking of conceptual content that allows us to understand how concepts can be applied in error within an individualist framework. There are certain cases of apparent ‘conceptual error’, however, that cannot be explained by appealing to facts about speakers taken individually. In these cases I argue that the shared language we speak determines the content of our concepts. This raises the possibility that there might be such things as ‘public concepts’ that are distinct from individual or private concepts, and to which we ‘defer’ in thought.

In chapter five I explore the notion of conceptual and linguistic deference, whereby we allow the opinions of others in our community to determine the meaning of some of our utterances or the content of some of the terms in our thoughts. For example, we might accept that the knowledge of the medical community determines the meaning of my uses of the term ‘arthritis’, or that the knowledge of the scientific community determines the meaning of my uses of ‘electron’. When I use these terms, I defer to the opinion of more knowledgeable members of my community on the meaning of the terms. However it is not clear that all cases of deference culminate in the opinion of an expert. I might accept that my uses of the word ‘shoe’ are deferential for the simple reason that I accept that I could be wrong about what the word ‘shoe’ means. But in such cases it is far from clear that there are experts in my community who determine the meanings of these words. Reviewing two problematic accounts of what might determine the ‘standing
meanings’ of terms in my language, I turn to recent literature on collective knowledge and integrate this with the notion of communicative value developed in earlier chapters to establish a model for what I call ‘participatory deference’. This model holds that the meanings of terms in shared languages can be modeled as functions of the beliefs of whole communities of speakers. Equally, I propose, this provides us with the material to make sense of the notion of a public concept that I discussed in chapter 4.

In the last chapter, I present an overview of the semantic model that has emerged. I argue that we can best think of this model as a ‘three-dimensional’ model of content – which adds to the two-dimensional model a third ‘public’ dimension. I end with some reflections on the implications of this view for the model of philosophical method as introspection that we have inherited from Descartes, arguing that we would be better off to think of much of the work of conceptual analysis as a political rather than a private project.
Reference in Thought

If I think about cigars, something in my thought refers to cigars. If I think about a piano, something in my thought picks out a piano. The terms in our thoughts – our concepts – somehow manage to pick things out in the world, or refer to them. In this chapter I review the major theories that have been developed to account for this phenomenon, and the problems they encounter. I ultimately endorse a version of what is known as ‘two-dimensional semantics’, which takes descriptions in the mind of a thinker along with facts about her mind-independent environment to determine the referential content of her concepts. Having laid out the full resources of the currently dominant theories of content, my aim in subsequent chapters is to identify the points at which these models fall short in providing an account of cognitive and linguistic phenomena, and to propose innovations to handle such cases. I begin, then, with the point of origin of much of the recent debates about content – Frege’s (1892) critique of Mill’s (1843) theory of meaning.

1. Content, Reason, and Modality

Intuitively, we might be inclined to say that the meaning of the word ‘water’ is just water. Similarly, we might be inclined to say that the content of my WATER concept is

\footnote{This first chapter is largely expository, so readers who are already familiar with the debates between descriptivism and the causal theory might want to skip to chapter two.}

\footnote{Throughout, I will use italics to indicate the objects/meanings/referents of referring terms (e.g. water), single inverted commas to indicate linguistic referring terms (e.g. ‘water’), and small caps to indicate}
the thing that my concept picks out: again, *water*. This intuition is the basic claim behind Mill’s (1843) simple model of conceptual content – the content of a concept is the thing to which it refers, and the meaning of a word is the thing the word refers to. Frege (1892) presents a powerful puzzle for this simple model. Suppose you know of two individuals, Dr. Jekyll and Mr. Hyde. As in the famous story, unbeknownst to you, Dr. Jekyll and Mr. Hyde are in fact two names for the same individual. You finally figure this out, and the following thought occurs to you:

1) Dr. Jekyll is Mr. Hyde!

Frege noticed that we can learn from 1) that the contribution made by the terms ‘Dr. Jekyll’ and ‘Mr. Hyde’ to a thought containing these terms cannot simply be the thing they refer to. If the only contribution those terms made to 1) was to refer to their bearer, then since the terms refer to the same bearer, the thought in 1) ought to be equivalent to:

2) Dr. Jekyll is Dr. Jekyll

However, the thoughts captured by 1) and 2) are clearly different thoughts. 2), after all, is knowable a priori – that is, knowable just by thinking about it, whereas 1) is not something you could figure out a priori.

To explain this phenomenon, Frege suggested that these terms were cognitively equivalent to descriptions, which he called the ‘senses’ of the terms. The term ‘Dr. Jekyll’ might be equivalent to the description ‘the man who tips his hat in the mornings’, and the term ‘Mr. Hyde’ with ‘the man to be avoided in the evenings’. In that case the thought in 1) would be equivalent to:

---

conceptual referring terms (e.g. *water*).

3 Robert Louis Stevenson’s (1886) “The Strange Case of Dr. Jekyll and Mr. Hyde”, published six years before Frege’s “On Sense and Reference”.

4 Frege presented his discussion in as a discussion of the meaning of words. However, it’s pretty clear he thought that the meaning of words is fully determined by the conceptual content of the speakers of those words, since he proposed to analyze the meaning of words in terms of the role they can play in thought. Hence, his discussion is primarily a discussion of thought content and only derivatively of linguistic content. I don’t agree that linguistic meaning is determined fully by the thoughts of any individual speaker, as I discuss in chapters 2-5. So for now I will present Frege’s view as a theory of conceptual content and avoid conflating these claims with claims about linguistic meaning.
3) The man who tips his hat in the mornings is the man to be avoided in the evenings

Since it is possible that an individual who satisfies the description ‘the man who tips his hat in the mornings’ might not also satisfy the description ‘the man to be avoided in the evenings’, it is clear that we could not know that 3) is true just by thinking about it. If 3) is an analysis of 1), in that case, we would have an explanation for why 1) could not be known a priori. We can call Frege’s proposal, that referring terms are cognitively equivalent to definite descriptions, ‘classical descriptivism’ (see also Russell, 1905; 1912).

The ‘descriptivist’ theory holds other advantages. If a name is cognitively equivalent to a uniquely identifying definite description, this will allow us to say exactly what a thinker’s thoughts are about. It appeared to Russell (1912) to be a prerequisite on attributing a thought about a particular object to an individual that the individual be in possession of knowledge of some criterion that would distinguish that object from all other objects in the universe. Without this knowledge, Russell supposed, there would be no way to decide which object the individual was thinking of. Following Byrne and Pryor (2006, 39) we can call this the ‘reference fixing role’ of a description.

Finally, the role of definite descriptions in our thoughts might explain how we are able to engage in ‘a priori’ reasoning. This is the form of reasoning that involves running through problems in our minds without appealing to empirical studies of the world – the archetypal example of which is mathematics. If our concepts are equivalent to descriptions of the objects that they pick out, we would have something to think ‘about’ in a priori reasoning. For example, if my thoughts about horses all involve some true descriptive information about horses, such as ‘four legged hoofed creatures with manes’, then this will explain how I can imagine a world that contains horses – by imagining a world that contains four legged hoofed creatures with manes (cf. Jackson, 1998a, 212; Chalmers, 2002b, 146). I might reason a priori, for example, that a world filled with horses would be different to a world filled with sheep: one would have whinnies and

---

5 For discussion see Evans (1982, 89ff).
snorts, the other would have wool. If I didn’t have any information about the things my concepts refer to, however, such a priori reasoning would not be easily explained.

Some don’t find this argument from a priori reasoning convincing (Byrne and Pryor, 2006; Soames, 2002, 2005), since it is not obvious how we could prove that we do in fact pick out the same objects in imaginary scenarios as our thoughts in the real world pick out. But there is another argument given in Evans (1982) that further bolsters the need for a priori content. Evans holds (following Strawson, 1959, 99) that attributing a thought about a particular object to a thinker must entail attributing the capacity to understand the difference between claims asserted to be true of that individual from similar claims made about another – his famous ‘Generality Constraint’. If a thinker possesses a concept of an individual – JOHN, for example – then the thinker must understand what it would mean for ‘John is happy’ to be true, and why this doesn’t necessarily mean the same thing as ‘Steve is happy’ (cf. Evans, 1982, 102). If we have no a priori knowledge of what distinguishes John from Steve, however, these thoughts ought to be cognitively equivalent to ‘someone is happy’, and hence to each other. And if a thinker is unable to evaluate claims about John in terms of John, and not someone else, then it seemed to Evans that this thinker ought not to be attributed a grasp of the concept JOHN. Together with the argument from the possibility of evaluating imaginary scenarios, we could call this the argument for a priori content from understanding.

These considerations have led many to endorse the claim that there must be an essential role for descriptive content to play in concepts, and that this role is essentially connected to the behavior of a concept in a priori reasoning. This intuition is expressed in what has been called the Fregean thesis about content (for two terms ‘A’ and ‘B’):

**Fregean Thesis:** ‘A’ and ‘B’ have the same sense if and only if ‘A=B’ is a priori

---

6 The theses of Frege, Kant and Carnap as constructed here are taken from Chalmers (2006, p.58), except that here I am articulating them over terms and not sentences.

7 ‘Iff’ abbreviates ‘if and only if’.

8 As Chalmers (2006) observes, Frege’s notion of cognitive insignificance is not in fact identical with the notion of a priority, since some mathematical truths are a priori but cognitively significant (a notion that Frege develops from Kant’s notion of the synthetic a priori). However, it is generally taken that the a priori and the notion of cognitive insignificance are very closely related if not identical, and so I will ignore this subtlety here.
The claim here is that if two concepts have the same sense, it would be a priori knowable that they are co-referring. Another way to put this is to say that terms will play exactly the same role in a priori reasoning just in case they have the same descriptive content.

In (1947) Carnap proposed that the notion of sense was essentially connected to another domain of reasoning – modal reasoning. This is the domain of thought that deals with the notions of necessity and contingency. As Carnap conceived of the relationship between necessity and contingency, necessary truths (such as $2+2=4$) are true in all possible worlds, or all ways the world might have turned out to be. Contingent truths (such as ‘it’s raining out’), on the other hand, are only true in some possible worlds, or some ways the world might have turned out.

To establish the connection between sense and modality, Carnap proposed that we begin by defining Frege’s notion of sense in terms of possible worlds. Carnap proposed that the sense of a concept can be understood as a connection that holds between the concept and a set of objects distributed across all possible worlds. A concept can pick out different objects in different possible worlds. For example, while the concept ‘the 44th US president’ picks out Barack Obama in this world, in another possible world it could pick out someone else. Since it is possible that Mitt Romney might have won the election, in another possible world the concept picks out Mitt Romney. The connection between the concept ‘the 44th US president’ and its many possible referents across possible worlds, then, is the sense of the concept. Carnap conceived of this as a mathematical function, which he called the ‘intension’ of the concept. He contrasts this with the ‘extension’ of the concept, being the object the concept picks out in any possible world. The sense, or the intension, is therefore a function from possible worlds to extensions – a set of connections between a concept and the many possible objects that could fall under that concept.

Having defined both notions of sense and modality in terms of possible worlds, Carnap was now in a position to clarify their relationship. First we can show that some laws can be established about senses, or intensions. If two terms have the same intension, they each will refer to the same thing as the other in every possible world in which they appear. For example, ‘female fox’ and ‘vixen’, have the same sense since they are both defined identically. As a result, ‘female fox’ and ‘vixen’ are necessarily co-referring in
any possible world. We can therefore say something about the relationship between necessity and intension – if two terms have the same intension, the objects they refer to are necessarily identical; vixens are necessarily female foxes. Carnap thought that we could go further than this, and say that if and only if two terms have the same intension, are the objects they refer to necessarily identical. This would give us the following strong definition of intensions in terms of necessity (for the terms ‘A’ and ‘B’):

**Carnapian Thesis:** ‘A’ and ‘B’ have the same intension iff ‘A=B’ is necessary

We now would have an elegant understanding of the relationship between meaning and modality – if two terms or concepts have the same intension, it is a necessary truth that the objects they pick out are identical; if two terms don’t have the same intension, the objects they refer to may be identical (as we saw above in the case of Dr. Jekyll and Mr. Hyde), but it is not a necessary truth since there are possible worlds in which those terms are not co-referring.

Chalmers (2006) shows us that at this point the prospect was on the table to add a third historic thesis about the relationship between modality and reasoning to these two to construct a ‘golden triangle’ inter-defining meaning, modality, and reason. This is Kant’s (1794) claim that all necessary truths were knowable a priori, such that if a claim was necessarily true, you could know it just by thinking:

**Kantian Thesis:** ‘A=B’ is necessary iff it is a priori

Adding this thesis to the first two would allow us to present an illuminating inter-definition of fundamental features of meaning (intensions and extensions) modality (necessity and contingency), and reason (the a priori and the a posteriori) as follows:

---

*Chalmers (2006) attributes the claim to Kant, although this proposal should perhaps be more correctly attributed to Hume (1740).*
Neo-Fregean Thesis:\(^\text{10}\) ‘A=B’ is a priori iff ‘A’ and ‘B’ are co-intensional iff ‘A=B’ is necessary.

This framework would have potent explanatory value. It would give explanations to the phenomena discussed above: the distinct roles of different but co-referring names in thought, the fixing of reference, and our ability to reason a priori. And the explanation that it gives has far-reaching implications. It predicts something about the nature of thought: that concepts must be constituted somehow by chunks of information about the things they are concepts of. And it predicts too that our rational capacities are ‘locked in’ to the nature of metaphysical reality – the limits of possibility – such that we can tell what is possible just by thinking about it\(^\text{11}\). As we shall see next, however, the role of ‘senses’ in this schema came under attack in a series of papers that were at the time thought to have wholly undermined the feasibility of the project.

2. Problems for Descriptivism and the Causal Alternative

Building on the work of Searle (1958, 1969), Kripke (1980) presented two serious problems for this framework, which we can call the ‘epistemic’ and ‘modal’ problems. Both of these undermine the claim that referring terms are cognitively equivalent to definite descriptions – a claim that, as we saw, Frege’s thesis relies on. The first is that it is not obvious that there is any definite description or group of descriptions\(^\text{12}\) that behaves

---
\(^{10}\) This formulation of the Neo-Fregean theses is entailed by Chalmers’ (2006) construal: “Two expressions ‘A’ and ‘B’ have the same intension iff ‘A=B’ is a priori’”; in the formulation presented here, I am restricting the discussion to referential terms not sentences, and explicating the role of all three of meaning, modality and reason in the formulation.

\(^{11}\) This last prediction is a staple of scholastic philosophy, put to famous use by Descartes who argued: if something is conceivable, it is possible; it is conceivable that my mind is not my body, so it must be possible that my mind is not my body; if my mind were my body, it would not be possible for my mind not to be my body since identity relations are necessary; therefore my mind is not my body (Descartes, 1641). The significance of the current issues for Descartes’ argument has been discussed extensively, but I will not explore that question further here.

\(^{12}\) Searle (1958, 1969) proposed that rather than abandoning the descriptivist theory and various advantages in light of these problems, we modify the theory such that referring terms are cognitively equivalent not to single descriptions, but to the disjunction of a collection of descriptions – a proposal that has been called the ‘cluster of descriptions’ theory. On this proposal, for example, we can doubt of single descriptions of
epistemically in the same way as a proper name. This means that we could stand in a
different knowledge relation to a proper name and the description it is supposed to be
equivalent to. We can illustrate Kripke’s claim by considering how it is possible to
entertain doubts about the connection between a proper name and any description it
might be thought to be equivalent to. For example, I could entertain the thought:

4) Perhaps Dr. Jekyll is not the man who tips his hat in the mornings

I can imagine, for example, that I mistook Dr. Jekyll for the man who tips his hat in the
mornings. Maybe I’m wrong about Dr. Jekyll being a hat-tipper, I can suppose. If,
however, the term ‘Dr. Jekyll’ was equivalent in my thoughts to the description ‘the man
who tips his hat in the mornings’, then 4) would amount to:

5) Perhaps the man who tips his hat in the mornings is not the man who tips his hat
in the mornings

Since 5) is incoherent, but 4) is not, it seems that the name ‘Dr. Jekyll’ cannot be
cognitively equivalent to this description after all. In contexts where I think about what I
know about the reference of my thoughts – ‘epistemic’ contexts – descriptions and proper
names do not seem to behave in the same way.

The term ‘Dr. Jekyll’ and the description ‘the man who tips his hat in the
mornings’ also seem to sometimes behave differently in modal thoughts – that is,
thoughts about other ways the world might have been. For example, the following claim
is clearly true:

6) Dr. Jekyll could have failed to be the man who tips his hat in the mornings

[Note: Jekyll whether they are true, while continuing to think of Jekyll under a disjunction of other descriptions, at
least one of which must be true of Jekyll. So we could doubt of Jekyll whether he is the man who tips his
hat in the morning, while continuing to think of him under a disjunction of descriptions such as ‘either [the
man who tips his hat in the morning, or the man I heard of from my neighbor, or the man whose name
appears on the cover of that book… etc.]’. However, in Kripke’s hands, it became clear that this solution
would not work for the puzzles Searle had identified: just as we can doubt whether any particular thing we
believe we know of the bearer of a name is true, we can doubt whether anything we know of the bearer of a
name is true.]
It is perfectly conceivable that the world could have turned out such that Dr. Jekyll was neither a hat-wearer, nor a hat-tipper. But if we substitute ‘Dr. Jekyll’ for ‘the man who tips his hat in the mornings’ in 6), we will get the obviously false:

7) Dr. Jekyll could have failed to be Dr. Jekyll

Supposing Dr. Jekyll had existed, whether he tipped his hat or not would appear to be contingent – he could have failed to do any of these things. But there is no possible scenario under which Dr. Jekyll could have failed to be himself. For this reason, Kripke argued that the cognitive value of proper names and natural kind terms is not identical to that of definite descriptions. He proposed to mark this distinction by calling proper names and natural kind terms ‘rigid designators’. Proper names and natural kind terms refer to the same thing, Kripke argued, across all possible worlds – while definite descriptions refer to different things across different possible worlds.

These observations caused two-fold damage for the Neo-Fregean model we considered above. First, the observations have important implications for the nature of the intensions of proper names and natural kind terms. If a proper name picks out the same individual in every possible world, then the intension of the name (the function that determines what the name picks out) will be identical for both the names ‘Jekyll’ and ‘Hyde’. Carnap’s thesis, that if two names have the same intension the identity of their bearers is a necessary truth, holds up under this observation. But Frege’s Thesis does not: since it is not a priori knowable that Jekyll is Hyde, it is not a priori knowable that terms with the same intension are co-referring. And entailed in this falsification of Frege’s Thesis is a falsification of Kant’s Thesis. Since Jekyll and Hyde are the same person in every possible world, it is a necessary truth that Jekyll is Hyde. But it is not knowable a priori. Therefore, all necessary truths are not knowable a priori.

*The Causal Theory of Reference*
Having seriously damaged the credibility of the role of descriptions in a theory of reference, Kripke was now left to explain what might take their place. The role of descriptions had explained three things, as discussed above: a) they provided material to play the ‘reference-fixing role’; b) they explained the cognitive significance of co-referring identities; and c) they provided material to play an ‘a priori’ role in thought, or to satisfy the ‘argument from understanding’.

To explain a), how the reference of terms is fixed, Kripke appealed to the causal connections between a thinker and the reference of the thought:

Someone, let’s say, a baby, is born; his parents call him by a certain name. They talk about him to their friends. Other people meet him. Through various sorts of talk the name is spread from link to link as if by a chain. A speaker who is at the far end of this chain, who has heard about, say Richard Feynman, in the market place or elsewhere, may be referring to Richard Feynman even though he can’t remember from whom he first heard of Feynman or from whom he ever heard of Feynman. He knows that Feynman is a famous physicist. A certain passage of communication reaching ultimately to the man himself does reach the speaker. He is then referring to Feynman even thought he can’t identify him uniquely (Kripke, 1980, 91)

Here Kripke outlines a scenario in which a person can refer to Richard Feynman apparently without having any uniquely identifying beliefs about Feynman, thereby putatively establishing a ‘reference fixing’ mechanism without the need for definite descriptions. Kripke’s model is also to be applied to natural kind terms – terms like ‘gold’, ‘water’, or ‘milk’. The term ‘gold’ will acquire its reference through a process whereby someone points at gold, says (or thinks) “gold”, and from that point on the term will pick out the substance that the original speaker was in causal proximity to.

13 Similar arguments are given in Putnam around the same time (see Putnam, 1970, 1973, 1975); and in Kaplan (1978, 1979, 1989) and Perry (1977, 1979). The latter arguments are based on observations about the functions associated with terms like I, now, today, here, actually and that. For example, Rip Van Winkle (Perry, 1977) wakes up in 1823 after sleeping for 20 years, and says ‘today is October 03 1803’. The description is false, of course, since the term ‘today’ has picked out the 3rd of October 1823 – the day on which it was uttered. In this case the description associated with the term does not pick out the reference. Similarly, Kaplan gives an example concerning the term ‘I’: Castor and Pollux are identical twins raised in qualitatively identical environments; the descriptions that either Castor or Pollux might use of themselves are accordingly identical; nevertheless, when Castor says or thinks ‘I’ the term refers to Castor, while in Pollux’s case, the term refers to Pollux. Here we have identical descriptions associated with utterances or thoughts that nevertheless have distinct referential contents, and the referential content cannot be determined by descriptions in these cases.
The causal theory therefore may suggest a possible answer for the first of the three questions we identified above, in a way that does not succumb to the epistemic and modal problems of descriptivism\(^\text{14}\). However it performs famously badly when attempting to answer the other two. Although several attempts have been made to resolve the puzzle of the cognitive significance of co-referring identities, none have been obviously successful\(^\text{15}\), and the importance of the a priori role seems not to have been recognized by the early proponents of the causal theory, since a priori content is wholly rejected by the causal theory. And so it may seem that the two competing theories are at a stalemate. One explains modal thoughts and cases of epistemic ignorance well, while the other explains the cognitive significance of co-referring identities, and provides traction for an account of understanding via a priori content.

A concise way of representing the conflict between the two models is as follows. The descriptive theory does a good job of explaining how the following thought can appear to be true:

8) Water might not be H\(_2\)O

As in the case of Jekyll and Hyde, if each referring term is cognitively equivalent to a description, there are some possible worlds in which water is not H\(_2\)O, and so the sentence is true. Entailed in the descriptive explanation of this phenomenon is an explanation for a priori reasoning, and material for reference-fixing. The causal theory does not explain well how a thought like this can seem to be true, given that the cognitive value of a term on the causal theory is just the thing the term refers to. On the other hand, what the causal theory does explain well is how the following thought, which seems to contradict the first one, also appears to be true:

9) Water couldn’t have failed to be H\(_2\)O

\(^{14}\) Although it faces a serious problem on this front – discussed in chapter 2 – known as the \textit{qua} problem.

\(^{15}\) There have been various proposals for a causal solution to these puzzles (e.g. Salmon, 1986; Braun, 1998, 2002; Richard, 1990; Ludlow and Larson, 1990; Crimmins, 1992; Soames, 2002, 2005), but none have secured a clearly satisfying account. Scott Soames, one of the chief defenders of the causal theory, ends one recent discussion of these issues with the admission that ‘the traditional problems for directly referential accounts of names have not gone away’ (Soames, 2005, 424).
If water is identical with $\text{H}_2\text{O}$, then there is no possible world in which water is not $\text{H}_2\text{O}$. Unlike the descriptive theory which predicts that 9) is false, the causal theory explains 9) well, because it takes both ‘water’ and ‘$\text{H}_2\text{O}$’ to have the same intension, or to be identical in all possible worlds. What we need a theory to do at this point is explain how both 8) and 9) can appear to be true in different contexts. Before turning to some options on that front, however, I will discuss next how descriptivism was not dismantled entirely by Kripke’s arguments. The promise of grounding reference in descriptions remained.

3. Reference-Fixing Descriptions

Following Kripke’s critique, various proposals emerged for how descriptivism might avoid the epistemic and modal problems. Here is Lewis on the prospects of a descriptivism that avoids the epistemic problem:

> What was well and truly defeated was a version of descriptivism in which the descriptive senses were supposed to be a matter of famous deeds and other distinctive peculiarities. A better version survives the attack: *causal descriptivism*. The descriptive sense associated with a name might for instance be *the place I have heard of under the name ‘Taromeo’*, or maybe *the causal source of this token: ‘Taromeo’*, and for an account of the relation being invoked here, just consult the writings of causal theorists of reference (Lewis, 1997, fn.22)

Here the suggestion is that we write the causal relation underlying naming into the description that is being used to ground the content of the term, in order to avoid Kripke’s epistemic puzzle. A description like ‘the man who tips his hat in the morning’ might fail to ground the name ‘Jekyll’ given that it is coherent to suppose that I am mistaken about Jekyll’s morning habits, but the description ‘the person I have heard of under the name ‘Jekyll’’ seems less susceptible to that critique – I can’t doubt whether the man I have heard of under the name ‘Jekyll’ satisfies the description ‘the man I have heard of under the name ‘Jekyll’’ (cf. Stalnaker, 1978; Kroon, 1997; Bach, 1994). There seem to be descriptions available of the objects of my thoughts, then, that are not susceptible to the
epistemic problem. Such descriptions would also explain the cognitive significance of co-referring identities – the descriptions ‘the person I have heard of under the name ‘Jekyll’’ and ‘the person I have heard of under the name ‘Hyde’’ are not a priori co-referring – the world could have turned out such that the person called ‘Jekyll’ was not also called ‘Hyde’. And so what we might call ‘nominal causal descriptivism’ – descriptions which appeal to the causal relationship between a speaker and the naming tradition upon which the reference of his thoughts and utterances depends – may be able to bridge the gap between descriptivism and the causal theory.

However it is not clear that such ‘nominal causal descriptions’ can play the reference-fixing role. It depends, after all, on a naming tradition already being place in my community to which I appeal when thinking of some person or object. Searle argued against this strategy early on however, say that while ‘reference to an individual may be parasitic on someone else’s [reference,] this parasitism cannot go on indefinitely if there is to be any reference at all’ (Searle, 1969, 170). In order to play the reference-fixing role, the description of the reference available to at least some thinker\(^\text{16}\) must distinguish the reference from all other things without depending on a prior naming tradition.

Searle (1983) proposed a candidate that might do this, while still avoiding the epistemic problem. Suppose I am currently faced with a blue object, and I want to give it a name. Well, it seems that I cannot be wrong that the object before me, whatever it is, is ‘the thing causing in me an experience of seeing a blue object’. Searle’s insight was that if we restrict the terms in our descriptions to phenomenal terms, we cannot be wrong that they apply to the object we are referring to. The reason is that phenomenal terms are, we might say, ‘incorrigible’: if I judge that something looks blue to me, then this is all it takes for it to be true that the object looks blue to me. Such judgments can be understood to be ‘immune to error’\(^\text{17}\). If I now index this phenomenal description to a time and place

\(^{16}\)In chapter two I am going to argue that a certain aspect of linguistic meaning – the communicative value of terms – is not reducible to individuals’ capacity to identify the reference of the term. But for now we are talking about how referring terms might be grounded for individual thinkers so that those terms take on particular meanings as they appear in the thinker’s own thoughts.

\(^{17}\)This phrase derives from its close cousin, ‘immunity to error through misidentification’; the latter made its way from Wittgenstein’s (1958, 66-7) discussion of how I cannot be mistaken about who the term ‘I’ applies to when I use it myself, through Strawson (1966, 164-5) to Shoemaker (1968). Evans (1982, 216, fn.23) points out that there is no need for the category of judgments to be restricted to ‘I’ thoughts, but includes a broader category of judgments, including any that do not depend on a correct identification of an object to which the predicate is being applied for their truth. Since ‘looks red to me’ judgments will never
(cf. Evans, 1979, 169), we surely have a reference-fixing definite description that cannot be false of the reference of my thought. I cannot, it would seem, be wrong that the object before me is ‘the object before me here and now that appears blue’.

While these phenomenal descriptions seems sufficient to play the reference-fixing role while handling the epistemic problem, a possibility that Kripke recognized, it is not clear that they can handle the modal problem\(^\text{18}\). The thought ‘Tarameo could have failed to be the place I have heard of under the name ‘Tarameo’’ is perfectly coherent, since the name ‘Tarameo’ could have been given to Rome. There is a possible world, in that case, in which ‘the place I heard of under the name ‘Tarameo’’ is Rome. It is not possible, on the other hand, that ‘Tarameo could have failed to be Tarameo’. So the referring term ‘Tarameo’ and the description ‘the place I have heard of under the name ‘Tarameo’’ still seem logically and cognitively distinct. But others have seen a way out of this problem. Evans (1979) proposed that we can ‘rigidify’ descriptions so that they would behave in the same way as rigid designators, and thereby avoid the modal problem as well as the epistemic problem\(^\text{19,20,21}\). To do this we modify the definite description ‘the man who tips his hat in the morning’, for example, with what has been called an ‘actually’ operator’ (Davies and Humberstone, 1980), which selects from among the multiple possible individuals who might have satisfied our description the one that satisfies the description depend on such an identification, they form a natural class with these others.

\(^\text{18}\) Cf. his discussion of the ‘descriptive names’ ‘Jack the Ripper’, stipulated to be whoever carried out the Whitechapel murders (Kripke, 1980, 79, fn.33) and ‘Neptune’ stipulated by Leverrier to refer to whatever was causing the discrepancies in the orbits of Venus. Kripke argues that even in such cases, for example that of Leverrier, no a priori knowledge of Neptune is thereby granted to Leverrier. Leverrier now knows that Neptune exists, for example, but since the name ‘Neptune’ has now been ‘rigidly’ assigned to Neptune, it is not a necessary truth that Neptune exists. In other words, Kripke thought that while descriptions could in fact avoid the epistemic problem, they wouldn’t avoid the modal problem.

\(^\text{19}\) Evans (1979) also argues that while it is possible to rigidify definite descriptions so that they avoid the modal problem, the ‘rigidification’ demands the inclusion of time and place operators (and maybe in some cases ‘ordering’ operators to handle descriptions with ambiguous scope, such as ‘the father of every girl is good to her’). Ordinary proper names and kind terms will not demand these modifications and so might not form a ‘natural kind’ with rigidified definite descriptions. However, Evans overlooks the fact that proper names themselves have to be rigidified to particular naming traditions that occur at particular places and times. If more than one naming tradition for the name ‘John Wayne’ exists, then the thought ‘John Wayne was tough’ will not unambiguously pick out John Wayne without specifying the naming tradition to which the name belongs, and this presumably could only be done via a description with time and place operators. For example, the thought ‘John Wayne is not John Wayne’ is not necessarily false, given that there could be two distinct naming traditions from which the name in the first and second instances derive their reference.

\(^\text{20}\) See also Kaplan (1979).

\(^\text{21}\) Another strategy for dealing with the modal problem is to reject that proper names are rigid designators (Dummett, 1973, 110-151); I won’t deal with that proposal here since I think the alternative is successful.
in the actual world. Instead of using a description like ‘the place I have heard of under the name ‘Tarameo’’, then, we appeal instead to a description like ‘the actual place I have heard of under the name ‘Tarameo’’, or ‘the place I have heard of under the name ‘Tarameo’ in the actual world’. Even were this expression to be evaluated in another possible world, the proposal holds, it would still refer to the place that is called ‘Tarameo’ in the actual world. We would introduce that operator into our description of Jekyll as follows:

10) Jekyll might not have been the actual man who tips his hat in the mornings

Since the ‘rigidified’ description picks out Jekyll himself, 10) amounts to the supposition that Jekyll might not have been himself, just as the same claim would with ‘Jekyll’ in the place of the rigidified description. And so the description ‘the actual man who tips his hat in the mornings’, assuming this is Jekyll, appears to behave just as the proper name ‘Jekyll’ itself does.

And so it seems like we have candidate descriptions that will pass both the epistemic and modal tests after all. In general we will have a canonical ‘reference fixing description’ available to appeal to that will adequately identify the reference of any concept while avoiding the epistemic and modal problems for earlier versions of description. For a concept like water, such a description might be ‘the actual stuff causing my experience as of watery looking stuff around here now’, containing a time index, a place index, and some phenomenal content adequate to identify the reference from among the various objects in a thinker’s environment, and an explicit appeal to the causal role the object plays in the thinker’s experience.

However, we’re not out of the woods yet. If the descriptions that are supposed to ground reference are rigidified using an ‘actually’ operator to handle the modal problem, then we lose the answer we appeared to have for why the first of our target sentences is true. Under a rigidified description, water is given an intension that picks out H₂O in every possible world, so that we can no longer explain the thought ‘water might not be

---

22 Here I will follow Chalmers (2006) in using ‘watery looking’ as a heuristic designed to capture the phenomenal content of an encounter with water.
H₂O’ as being true because water is not H₂O in some possible world. On the other hand, if we don’t rigidify the descriptions, then we have no explanation for why it is a necessary truth that water is H₂O – since an ‘unrigidified’ description will not pick out H₂O in every world. We need a model that shows why both are true at once.

4. Two-Dimensionalism and Complex Intensions

A newer development in this area is known as two-dimensional semantics. There are various versions of two-dimensionalism²³, and in each it is proposed that there is more than one intensional function at work in determining the truth-values of our target sentences. A model with particularly strong currency is that of David Chalmers, which I will call the ‘complex intensions’ view. But first some background on the two-dimensional proposal.

The key innovation of the two-dimensional proposal is to recognize that the modal operators in the target sentences 8) and 9) – the terms ‘might be’ and ‘couldn’t have failed to be’ – are ambiguous. In the first sentence, it is argued, the term ‘might’ captures what have been called ‘epistemic possibilities’²⁴ for the reference. What a given reference ‘epistemically-might’ be is determined by what the person thinking about that reference knows about it. This means that anything is epistemically possible that doesn’t conflict with what I already know. What I already know about the reference is going to be fixed by the reference-fixing description that is used to determine the reference of the term. So, in the case of a term fixed by the description ‘watery-looking stuff’ (along with the various indices discussed above), what I know a priori is that the reference is watery-looking. Since I do not know a priori that anything that is watery-looking is H₂O, knowing that the object is watery looking does not rule out the possibility that the it is

---

something other than H₂O. We can then say that it is ‘epistemically possible’ that water is not H₂O.

In the second sentence, on the other hand, the modal operator (in this case ‘couldn’t have failed to be’) captures what have been called ‘metaphysical possibilities’ of the reference – what the reference could be given the state of the actual world, including the actual state of water, and irrespective of my current state of knowledge of the reference. So, for example, water couldn’t have failed to be H₂O since it actually is H₂O. There is no possible world in which water isn’t itself. Since water is H₂O, and identity is a necessary relation, it is not metaphysically possible that water could have failed to be H₂O. Since one type of possibility depends on what I know, but the other doesn’t, some particular claim can be simultaneously epistemically possible and metaphysically impossible. We now begin to get a glimpse of how the sentences ‘water might not be H₂O’ and ‘water couldn’t fail to be H₂O’ can be true at the same time. If the first is entertained by a thinker who does not know that H₂O is water, it can be true, since its truth is determined by the epistemic state of that thinker; the second can be true independently of what the thinker knows, and hence at the same time as the first.

Chalmers (2006) proposes that we can use these different aspects of modality to provide an ‘intensional’ account of the two target sentences, and hence a descriptivist account of reference that can be explained in terms of possible worlds – potentially reconstructing the ‘golden triangle’ that seemed to be within the reach of Carnap. He proposes that every term has two separate intensions, one corresponding to the epistemic domain and the other to the metaphysical, and depending on what sort of thought the term appears in, one or the other intension for the term will be used to evaluate the truth of the thought. In thoughts like ‘water might not be H₂O’, truth is evaluated against the term’s ‘epistemic intension’²⁵. The epistemic intension of the term is like a non-rigidified description – it picks out anything in any possible world that, in this case, would satisfy the amount of descriptive information we currently have about the reference²⁶ – in this case simply that it is watery-looking. And so we will be given an explanation for why we say that it is true (if we don’t know otherwise) that ‘water might not be H₂O’: in some

²⁵ Later called a primary intension.
²⁶ Which Chalmers calls the terms ‘epistemic profile’.
possible worlds, there is watery-looking stuff that is not H\textsubscript{2}O. The epistemic intensions of ‘water’ and ‘H\textsubscript{2}O’ do not map onto one another exactly. On the other hand, in ‘subjunctive’ contexts of evaluation – where we think about how the actual things in the world could have turned out in other possible worlds – the truth of the claim is evaluated according to the term’s ‘metaphysical intension’\textsuperscript{27}. The content of the metaphysical intension is identified by taking the term as a rigid designator, allowing essential nature of the substance in relation to which the thinker of the term stands in a causal relation to determine the intension of the term. If the term ‘water’ is uttered on earth, then, its metaphysical intension picks out H\textsubscript{2}O in every possible world. And this explains why the claim ‘water couldn’t have failed to be H\textsubscript{2}O’ is true: there is no possible world where H\textsubscript{2}O is not H\textsubscript{2}O. The metaphysical intension of ‘water’ and ‘H\textsubscript{2}O’ are identical.

This gives us the contrasting truth conditions that we sought after for our two target sentences, explained now in terms of intensions. The first is true given that the ‘might’ in the sentence is given an epistemic reading, as in ‘might for all I know’:

\begin{quote}
8*) Water might [epistemically] not be H\textsubscript{2}O
\end{quote}

Since the ‘might’ takes the force of ‘for all I know’, the sentence is true just in case we are not in a position to tell whether the molecular structure underlying the phenomenal presentation of watery-looking stuff is H\textsubscript{2}O or not. The second sentence can be simultaneously true, even though it seems to contradict the first, given that the modal operator is taken in the metaphysical sense:

\begin{quote}
9*) Water couldn’t [metaphysically] fail to be H\textsubscript{2}O
\end{quote}

If we take the modal operator here to be metaphysical, it operates over the term ‘water’ taken in the sense of ‘whatever water actually is, irrespective of what I know about it’. That fixes the intension of the term as a function that picks out H\textsubscript{2}O in every possible world, such that 9* is true.

\textsuperscript{27} Also called ‘subjunctive intension’ and ‘secondary intension’.
We now have the option of restoring the ‘golden triangle’ inter-defining meaning, modality and reason. Although it is not true that two terms have the same epistemic intension if and only if it is a necessary truth that their references are identical, two terms do have the same metaphysical intension if and only if ‘A=B’ is true in all possible worlds. And so metaphysical intensions and necessity can be fully inter-defined. Turning to the relationship between reason and modality, it is not true, as Kripke showed, that if some claim ‘A=B’ is necessary, it will be knowable a priori. It is not true, then, that if two terms have the same metaphysical intension, the identity of their bearers will be knowable a priori. However, it is true that if and only if two terms have the same epistemic intension, the identity ‘A=B’ will be knowable a priori. And so we can fully inter-define the a priori and the epistemic dimension of meaning. We seem, therefore, to have a restoration of the ‘golden triangle’, and also a restoration of a systematic role for a priori content to play in the determination of meaning. Necessity can be defined in terms of the metaphysical dimension of meaning (the metaphysical intension), while the a priori can be defined in terms of the epistemic dimension of meaning.

We can now also explain the three criteria mentioned in section 1. The descriptive content of the epistemic dimension of meaning provide us with knowledge of the reference sufficient to distinguish it from every other object in the universe (Russell’s Criterion), explain the cognitive significance of co-referring identities, and provide us with a priori knowable information about the objects to which they apply, grounding our capacity for reasoning about the referents of those terms (Evans’ Generality Constraint).28

There is plenty of dissent to this approach (Hawthorne, 1996; Bealer, 2002; Schiffer, 2002; Marconi, 2005), but the dissent arises largely in the form of challenges to details of the view as opposed to discoveries that the view gets the wrong results (although see Soames 2002, 2005, 2006 for sustained arguments against the role of descriptions in grounding reference; see Chalmers 2006; 2007 for replies29). There is

---

28 What we have not established is whether Kant’s thesis is true in any way; but nothing hinges on that in the current work, so I will not try to decide the issue here.

29 For example, Soames (2003, 2005) rejects actually-rigified descriptions’ capacity to properly account for belief attribution. Soames presents his argument as a challenge to causally self-referential descriptions of the form proposed by Lewis (1997) and Searle (1983), but his argument applies to the current model also. His worry is that if the term ‘Venus’ in my thoughts is cognitively equivalent to a description like ‘the actual bright thing in the sky at time t, location l’, then it will not be possible for, say, ancient Babylonians to share with me any beliefs about Venus. Their beliefs couldn’t possibly refer to the time and location at
surely plenty more to learn about the nature of the relationship between epistemic and metaphysical intensions, but what I think is clear is that both descriptive a priori knowable, internal, or ‘representational’ content and also causal external or ‘relational’ content play a role in grounding the meaning of the terms in our thoughts. We have arrived, then, at the current state of the art of modeling conceptual content.

Conclusion

In spite of the misgivings of theorists at various points over the course of the development of recent theories of content and meaning, it seems that the prospect of a unified account of referential thoughts is before us after all. What I want to explore over the next five chapters are points at which this account is insufficient to handle the full range of cognitive and linguistic phenomena, and the ways in which it might be extended in order to handle some puzzling cases. Particularly, I hope to show that the meanings of terms in the languages we use are underdetermined by the full resources of a two-dimensional model, and hence are underdetermined by the cognitive content of any individual thinker or speaker. In chapter two I turn to a respect in which the two-dimensional model underdetermines the content of our concepts and the meanings of new terms introduced to a language, in a puzzle known as the ‘qua problem’.

which I apprehend Venus, so we will not be entertaining the same thought, and hence not sharing a belief that we plausibly will have shared. This is the same puzzle that Kripke (1979) identified for Pierre who simultaneously believes that London is not pretty and that Londres est Jolie. A solution to these cases I think is available by simply stating that it’s possible to attribute beliefs truth conditionally. That is, there is a sense in which Pierre both believes and disbelieves the same belief, and a sense in which Pierre has two distinct beliefs with conflicting truth conditions. It all depends how we individuate beliefs. If we say of the ancient Babylonians that they share beliefs with the same truth conditions as us, that may be all we need to satisfy the intuition that they share beliefs with us. Also see Chalmers (2007).
In the last chapter I proposed a model for the determination of the meaning of the terms in our thoughts that fixes the reference of those terms using descriptions in a way that avoids the epistemic and modal puzzles for classical descriptivism. A difficulty for this story now appears in a puzzle known as the ‘qua problem’ (Schwartz, 1978; Papineau, 1979; Dupre, 1981; Devitt, 1981; Sterelny, 1983; Devitt and Sterelny, 1987; Hilary Kornblith, 1980; Miller, 1992; Brown, 1998; Prinz, 2002; Jylkka, 2008). The problem is that the ostensive definitions by which we introduce new terms to a language do not seem to fully determine the reference of those terms. Many possible objects of reference occupy any ostensively identified region, and it is not clear what determines which possible object is being identified. The problem arises for both the content of demonstrative thoughts, and also for the meaning of newly defined linguistic terms.

1. The Qua Puzzle and the Causal Theory

As we saw in the last chapter, Kripke supposes that both proper names and kind terms acquire their meaning as a result of a causal chain of speakers that begins with an initial baptism, and ends with the utterance of the name. In the initial baptism of an individual, the baptizer points to the referent (a baby, for example, but also possibly a car, or a pet, etc.), and announces the newly introduced name. In the baptism of a natural kind like gold, the baptizer points to a sample of gold which he picks out by its phenomenal properties, announces the name, and forever after the name is fixed to the actual kind instantiated by this sample.
The *qua* problem subsequently appears as a puzzle for the determination in these acts of ‘ostensive definition’ of the meaning of the term thus defined. The problem is that, in the case of the introduction of a kind term, the sample that is used to define the kind term is likely to simultaneously embody *multiple* kinds. A sample of gold, for example, is simultaneously a sample of gold, a sample of metal, a sample of dense matter, a sample of a particular shape, and also it is a particular object and not just a sample of a kind – it is a spatio-temporal object, it also embodies a time slice of a spatio-temporal object, and so on. As Sterelny and Devitt present the puzzle:

The term is applied to the sample not only *qua* natural kind but also *qua* member of one particular natural kind. [However,] any sample of a natural kind is likely to be a sample of many natural kinds; for example, the sample is not only an echidna but also a monotreme, a mammal, a vertebrate and so on. In virtue of what is the grounding *qua* member of one natural kind and not another? As a result of groundings, a name refers to all objects having the same underlying nature as the objects in the sample. But which underlying nature? The sample shares many (Sterelny and Devitt, 1987, 73)

The act of ostensive reference – simply pointing at the sample and saying ‘gold’ or ‘echidna’, is somehow supposed to ensure that one and only one from among these possibly indefinitely many instantiated kinds and objects is identified as the referent of the kind term. The question is: what determines which instantiated object is named.

The same puzzle appears for the introduction of proper names. Feynman’s parents places their hands on their baby’s forehead, and announce ‘Richard Feynman!’; but why is it that the name comes to pick out Feynman the person, and not just Feynman the baby, or a time-slice of the baby, or the baby on a Tuesday? All of these things are in contact with his parents’ hands. An earlier discussion in Wittgenstein anticipates both the Kripkean proposal and the inevitable problem it will face:

[Perhaps] one can ostensively define a proper name, the name of a colour, the name of a material, a numeral, the name of a point on the compass and so on. The definition of the number two, “That is called ‘two’” – pointing to two nuts – is perfectly exact.

– But how can two be defined like that? The person one gives the definition to doesn’t know what one wants to call “two”; he will suppose that “two” is the name given to *this* group of nuts! [Or,] he might make the opposite mistake; when I want to assign a name to this group of nuts, he might understand it as a numeral. And he might equally well take the name of a person, of which I give an ostensive definition, as that of a color, a race, or even a point of the compass (Wittgenstein, 1953, §28)
It might appear that that the *qua* puzzle is distinctive to the causal model of reference, and that since I have already adopted a descriptivist model, raising the puzzle here is redundant. We shall see however that the problem arises for the descriptivist approach too.

### 2. The Qua Puzzle and the Descriptive Theory

Devitt and Sterelny (1987) suggest that the solution lies in including some descriptive terms in the thoughts of the baptiser. “It seems”, they suggested, “that the [baptiser] must at some level ‘think of’ the cause of his experience under some general categorical term like ‘animal’, or ‘material object’” (Devitt and Sterelny, 1987, 64-65). Here it is assumed by Devitt and Sterelny that if a small amount of descriptive content can be allowed into the baptismal act, then the problem will be resolved. Similarly, Putnam (1975, 269) proposed to avoid the problem in advance by including a ‘marker’ – when I baptize *water*, for example, I say something like “this liquid is called ‘water’” – the marker ‘liquid’ rules out other kinds instantiated by the sample.

The first problem for this approach is that it involves what we might call cheating. We can’t appeal to a stock of already determined semantic terms to solve the problem of how we are to ground the reference of terms of that very sort. Our challenge is to explain how the reference of a demonstrative thought is fixed as a member of one class and not another, or as a member of a kind versus an individual, and so appealing to other terms that have already had their reference fixed to a kind (such as ‘animal’) is presupposing that we have an answer to that question. How, for example, was the reference of the term ‘liquid’ that appears in the reference-grounding description fixed? Solutions of this sort therefore appear to summon a regress. But there is a deeper problem still for descriptivist solutions to the puzzle.

The deeper problem is that the epistemic puzzle that we considered in the last chapter crops up again, but in such a way that the solution we considered in the last chapter will not avoid it. In order to avoid the epistemic problem for descriptions that can
be used to ground reference, we had to restrict the terms in the description to immediately perceptible phenomenal terms. I can’t be wrong that something is currently ‘watery looking’ to me, because the condition for something to be currently watery-looking just is for me to believe that it is currently watery-looking. However, such incorrigible phenomenal descriptions will only suffice to narrow reference down to an area in the vicinity of the thinker. Since the different objects between which we are trying to discriminate are all co-instantiated in a single region in the vicinity of the thinker, however, they will all satisfy a phenomenal description that does no more than pick out that region. And so descriptions presented in these incorrigible phenomenal terms won’t discriminate between water, the particular volume of water before me, liquid, earth bound substance, and so on. Similarly, a phenomenal description of a sample of gold designed to fix the reference of ‘gold’ will not distinguish between “the isotope $^{197}$Au (the only naturally occurring isotope of gold), the element gold, transition metal, precious metal, etc.” (Jylkkä, 2008, 156).

The only descriptive terms that would distinguish between these possible referents will be kind terms, of the sort discussed above. But terms like this are not immediately verifiably true of the objects that I pick out in reference. The cats in my vicinity could turn out to be robots or the pencils tiny organisms (Putnam, 1975, 243). But I can surely have demonstrative thoughts about cats without knowing whether they are robots or animals. We might think we could avoid the problem, in that case, by identifying the reference of the description with an extremely general term like ‘thing’ or ‘stuff’. This is the form of the reference-fixing descriptions in Searle (1983), Jackson (1998a, 1998b) and Chalmers (1996, 2002, 2006) – Chalmers’ reference-fixing description for ‘water’ being ‘the watery looking stuff around here’. Terms like ‘thing’ or ‘stuff’, after all, cannot fail to be true of the reference of really any thought, and so will avoid the epistemic problem. However, extremely broadly applicable terms like this will not distinguish between the various possible objects of reference, for the simple reason that the term ‘stuff’ will be true of them all. This difficulty is expressed nicely as a dilemma in Miller (1992):

The more definite the … description is, the more vulnerable it is to counterexamples along the lines of Putnam's robot cats. The less definite the … description is, the less
helpful it is in eliminating unintended classes (Miller, 1992, 428)

Either the descriptions will include terms that are specific enough to distinguish between the various possible objects of reference instantiated in an identified spatio-temporal region, in which case they will not be incorrigible and will succumb to the epistemic puzzle for descriptions; or else they will be so general that they will fail to distinguish between the many possible referents.

**Recognitional Capacities and Transformation Laws**

An alternative to including kind terms in our descriptions is to appeal to *recognitional capacities*, made by theorists on both sides of the descriptive/causal divide. Sterelny (1983), Miller (1992), Brown (1998), Jackson (1998b), and Jylkka (2008) all propose that such an appeal will provide a solution to the *qua* problem. Sterelny, for example, suggests that the members of a particular kind will have causal powers that include the power to get us to recognize members of that very kind. Imagining the role recognitional capacities play in the introduction of a kind term for the novel species ‘schmats’, Sterelny argues:

> It is not just that the exobiologist acquires the ability to recognize schmats, rather than schmat-sexes, or the genus that schmats exemplify, though this, too, is relevant to its being schmats that are picked out. But further, the recognitional capacity itself is one instance of the causal powers of the kind. I can recognize schmats because of certain of their macroscopic features though I may not know which or how (Sterelny, 1983, 110)

Having the ability to recognize a member of a kind does not depend on having the ability to describe the kind, Sterelny suggests here. I might be able to recognize a Swede, perhaps because I have some Swedish friends, even though I might not be able to tell you what subtle characteristics of Swedes uniquely allows me to recognize them, just as Evans discusses is true in the case of our ability to recognize our friends (Evans, 1982, 286). This presents a clear advantage for the causal theorist, who wants to exclude descriptive content from the reference-fixing act, while also plausibly singling out the kind that the term refers to. The recognitional capacity that is triggered when I encounter
a kangaroo, for example, may be triggered by kangaroos but not by any other animal. This may appear to help to isolate the reference of my thought to a particular kind.

The descriptivist, on the other hand, who is not happy that causal relations are sufficient to establish the content of a referring term, may propose that a reference to the recognitional capacity currently deployed in my apprehension of a particular object can be included in the reference-fixing description. Brown (1998) and Jylkka (2008), for example, propose that a term can acquire a reference by a thinker associating a particular recognitional capacity with the term. This suggests that we build in a reference to a particular recognitional capacity into the description that fixes the reference. Perhaps the description would be something like ‘the thing that the recognitional capacity that is being currently triggered picks out’.

However, appealing to a triggered recognitional capacity in the description is problematic for two reasons. The first reason is that many recognitional capacities can be deployed at once – indeed, as many as there are possible objects of reference. So, in the apprehension of an eagle-chick I recognize an eagle, an eagle-chick, a time slice of an eagle-chick, an animal, and so on. The description ‘the recognitional capacity currently being deployed’ is therefore going to face the qua puzzle all over again. The second reason is that the transformations and variations that the recognitional capacity would permit, which are ‘implicated in the subpersonal level whirrings and grindings of the device that underpins my recognitional capacity’ (Davies, 2004, 166) may be completely out of reach of my personal level beliefs. This would mean that the objects of my thoughts would be fixed without my having any understanding of what sort of things those objects were. As we discussed in chapter 1, in order to attribute a thought about a particular object to a thinker we plausibly need to be able to attribute to the thinker the capacity to understand what it would mean for an indefinite number of claims to be true.

---

30 Brown (1998), p.284: “natural kind terms gain their reference not by being explicitly defined but rather by being associated in the way described with a recognitional capacity”; Jylkka (2008) in turn proposes that the reference of a term is fixed as the kind “whose instances trigger in the relevant speakers of C (in w) the recognitional component of the concept associated with t”. Jylkka goes on to include a more complex definition that takes into account differences between superficially recognizable kinds and kinds discriminated by their ‘deep’ or hidden properties, but the appeal to an associated recognitional capacity remains in the definition.

31 Which I owe to Bill Seager.

32 I owe the genesis of this second line of thought, along with my appreciation of the importance of personal level cognitive states to reasoning to Sonia Sedivy.
or false of that object (cf. Evans, 1982, 100), or to understand what it would mean for claims to be true of that object as opposed to some other. In the case of demonstrative identification, this means that the thinker needs to understand what sort of object he is identifying – and that it is not sufficient for him simply to appeal to ‘whatever sort of thing my recognitional capacity will track’.

This is a serious problem for the appeal to recognitional capacities. For example, Evans allows that we appeal to recognitional capacities to track objects in our environment as long as the capacity provides the thinker with an ‘unmediated disposition to treat information from that object as germane to the truth and falsity of … thoughts involving that [object]’ (ibid, 146). That is, he allows that if I can continue to track an object in space given an initial identification, then I have a discriminating conception of the object sufficient to ground reference. The disposition that Evans speaks of is a disposition to locate on object in space and track that location over time. His model thereby holds that ‘it follows that one can discover oneself to be radically mistaken about the object of one’s thought’ (Evans, 182, 179, my italics). This sets us on the right track to solving the problem, but doesn’t get us all the way. What it allows is that we can have personal level knowledge, at the time of identification, of what object we are identifying insofar it can be distinguished spatio-temporally. For example would be able to state given a little a priori reflection that whatever object we are currently identifying must remain causally continuous with the region before us. But we can see that this will not do for two reasons. First, Evans’ presupposes that there is just one object in any spatio-temporal region, as we can see from his recurring reference to ‘the object’ that we are in contact with. And secondly, of course, many objects can be co-instantiated in a single trackable chunk of material: time slices of eagles, eagle chicks, long-lived eagles, eagle-feather hats, and so on. If all we can say is that the object we are identifying must remain spatio-temporally continuous with the region we have identified, then it is still left indeterminate what we’re identifying. The qua problem returns with full force.

What we need is a way of incorporating the work that a recognitional capacity can do to distinguish between the possible objects of reference into our a priori knowledge of

---

33 Which Evans calls ‘public space’. The idea is that a real causal connection between one point in the external world and the subject is preserved.
the reference. To see if there is a way of doing this, first we have to state more precisely exactly what work the recognitional capacity is doing. The difference between what a recognitional capacity might ‘pick out’ and what a phenomenal description of a chunk of space-time could pick out, is that the recognitional capacity will continue to pick out a single object of reference even as it undergoes substantial changes in phenomenal appearance, or presents itself under radically different phenomenal appearances. The only way to distinguish between the possible objects instantiated in a time slice of a spatial region is over time, and counterfactually.

For example, the difference between an eagle and an eagle-chick that are both instantiated in one spatio-temporal region is that while the eagle chick will cease to exist after it turns into a fledgling, the eagle will persist through the transformation into a fledgling, and then into an adult. These transformations distinguish the eagle from the eagle chick, and so a mechanism that places constraints on the possible transformations that an object can undergo will distinguish between these two possible objects of reference, in a way that the ‘snapshot’ phenomenal description cannot. These possible transformations and variations in appearance have been explored under the notion of ‘Fundamental Ideas’ (Evans, 1982), ‘constitutive rules’ (Haugeland, 1994), ‘theories’ (Gopnik, 1997), and ‘Governing Conceptions’ (Dickie, 2009, 2010). For example, if a chair suddenly, somehow, evaporates, then most of us will not count the vapor as the same object as the chair – we are more likely to think the chair no longer existed. In that case, we might say that the law governing which object a recognitional capacity will trigger in response to will prohibit the recognitional capacity from continuing to trigger in response to vapor, if the recognitional capacity is for a chair. Similarly, our capacity to perceive an object as belonging to one kind and not another is often argued to be conditional on our possessing knowledge of what transformations in perception that object would undergo if it was an object of one type and not another (Cf. Baldwin, 2003; Hurley, 1998; Kelly, 2001; 2004; Thompson and Varela, 2001; Noë, 2004; Schellenberg, 34

---

34 Obviously I’m playing fast and loose with notions of subsistence here, but at least it should be clear that the baby ceases to exist insofar as the referring expression ‘the baby’ might apply to it; and certainly, the eagle chick will cease to exist when the eagle dies, while the material out of which the eagle chick is made at a certain point in time will not.
The rules governing what a recognitional capacity would trigger in response to would be sufficient, then, to distinguish among the various objects instantiated in any region of demonstrative identification. What we now want to ask is whether we can include the rules or governing laws that an object would have to ‘obey’ to count as one kind and not another in our reference-fixing description. The obvious problem here is the risk of falling back into the epistemic problem. If the object of reference turns out not to obey the governing laws the thinker expects, for example, then we might think that the governing laws could not play a role in fixing the reference. For example, if I point to what appears to be an animal in the dark, and say ‘that’s hungry’ (Dickie, 2010, 230), we might expect that our governing laws would demand that the object be at least an animal, so that if it turns out to be simply a shadow, reference will not succeed. However, I don’t think this is an option given the extent to which we could be wrong about the references of our thoughts. If the object to which we ascribed hypothetical hungriness turned out not to be an animal but a well-disguised robot, it becomes less plausible to say that we never referred to the object because we were wrong about it’s capacity for hunger, instead of saying that we falsely attributed the possibility of being hungry to a robot. We can approach the robot, pet it, see it wag its tail, and so on. It is unconvincing to suppose that we are not under such circumstances referring to the robot.

So our governing rules are going to have to accommodate the possibility that we have almost entirely false beliefs about what kind of thing it is to which we are referring, with the exception of it presenting the incorrigible phenomenal properties we perceive in it, and still succeed in isolating the reference of the demonstrative as one among the many objects instantiated in the demonstrated region. This is the joint challenge of the qua problem for thought.

A solution to this puzzle, I propose, is available in the vicinity of the idea of governing rules. We need to understand the governing rules as a set of conditionals. For example, if I see a shadow by the wall that I think might turn out to be a dog, or might

---

35 This raises the further question, debated in these papers, of whether knowledge of the transformations such objects could undergo is constitutively connected to certain other capacities such as the capacity to interact with those objects; I will not explore this question here, as what is currently at stake is not the question of ‘how’ I acquire such knowledge, but only whether it can be used to inform a reference-fixing description.
turn out to be simply a shadow, a conditional will govern the application of my demonstrative ‘that' like:

If that turns out to be a dog, then it will remain as the reference of my demonstrative until it dies

What this allows is that a) the reference is not required to be a dog to be the reference of my demonstrative; and b) the qua of the reference, if it is a dog, is isolated from among ‘dog’, ‘time slice of dog’, ‘material making up a dog’ and so on. Under a) it is allowed that even if I am wrong about the object being a dog, reference may still have succeeded. So if it is simply a statue, then an added conditional such as:

If that is a statue, then it will remain as the reference of my demonstrative until it is demolished

will allow that the demonstrative refers to the statue, and again distinguish the statue from its clay. And if it is simply a pattern of light on the wall, then a conditional like:

If that is a pattern of light, then it will remain as the reference of my demonstrative until it dissipates

will determine what relationship holds between my demonstrative and the multiple possible objects in the ostended region under the circumstance that there is nothing there but a trick of the light. Since all of these conditionals can hold at the same time, we now have the possibility of articulating governing laws that can accommodate indefinitely many possible discoveries about the object of thought, while at the same time constraining reference to a particular ‘qua’ depending on what the object turns out to be.\footnote{A very closely related proposal designed to allow for the perception of objects as belonging to particular categories is to be found in Campbell (2002), 237ff.}
What all this suggests is that in order to solve the qua problem for demonstrative thought, the conditional governing laws determining the object type must be written into the reference-fixing description\(^\text{37}\). The general form of reference-fixing descriptions for the content of terms fixed by demonstrative identification would therefore be:

The *phi* that obeys conditional governing laws *L*

Here ‘*phi*’ is the reference-fixing description that we discussed in chapter 1, which we have now supplemented with conditional governing laws to resolve the qua problem. And so we seem to have at least the promise of a solution to the qua puzzle for the content of demonstrative thought. However, while the literature on this topic assumes that a solution to the qua puzzle for demonstrative thought provides a solution to the *qua* problem for linguistic reference, I now want to show that such is not the case.

### 3. Why we still haven’t Solved the Puzzle

The difficulty faced by the solution considered above, for what determines the object of an individual speaker’s thoughts, is that it is not at all clear that such a story will be sufficient for grounding *linguistic* meaning. In addition to the question ‘what determines the object of the speaker’s thought’, we need to explain how his *utterance* ‘Aristotle’ (for example) takes on a meaning for the linguistic community to which he introduces it. There are, therefore, two distinct questions at work here. To return to Wittgenstein’s discussion of the puzzle, we can remind ourselves that part of the *qua* problem is concerns how a speaker’s *audience* can understand his meaning:

\(^{37}\) It might be worried that this suggests that indefinitely many ‘rules’ are going to have to consciously be endorsed in every demonstrative identification (cf. Kripke, 1982, 7-22); but this does not follow – the rules need only be ‘a priori available’, that is, available given a little a priori reflection. My ‘rational status’ at any time will surely include my endorsement of an indefinitely large set of beliefs, but I don’t need to repeat them to myself every time I undertake some project for which their endorsement is required.
The person one gives the definition to doesn’t know what one wants to call “two”; he will suppose that “two” is the name given to this group of nuts! (op. cit.)

To illustrate how the determination-of-baptizer’s-thoughts approach to the qua problem, fails to answer the question about linguistic meaning, let’s consider a hypothetical example of the introduction of a name into a language, and allow for the determination within the speaker’s thoughts of the object in question. We will see that it is far from obvious that the content of the speaker’s thought will be automatically extended to the speaker’s utterance. Returning to the baptism of Aristotle, let’s accept that the baptizer’s thoughts determinately pick out Aristotle the person – not as a time-slice of a person, or as a person until the year 2010 and Emerose thereafter, or as a baby, or a baby in a particular baptismal outfit, but as a person. Now consider the following cases:

Case 1: The priest intends to baptize Aristotle the person; the congregation believes the priest to have baptized the person

Case 2: The priest intends to baptize the Aristotle person; the congregation believes the priest to have baptized the baptismal frock

Case 3: The priest intends to baptize the baptismal frock; the congregation believes the priest to have baptized the person

Case 4: The priest intends to baptize Aristotle the person; the four members of the congregation (A, B, C and D) believe the priest to have baptized (A) Aristotle the person, (B) a 3 minute time-slice of the Aristotle beginning at the moment of baptism, (C) the baptismal frock, and (D) the point on the compass toward which the priest’s forefinger extends

Now let’s imagine that in each of these cases, the priest drops dead immediately after the baptizing ceremony, so that he cannot be subsequently consulted about his intentions. What will we say about the meaning acquired by the name ‘Aristotle’ in each of these cases?

In case 1, it seems like a naming tradition may have been successfully established. The priest’s intentions are understood by the congregation, and presumably the members of the congregation will ask each other questions about ‘Aristotle’, and get appropriate

responses, each one believing of the others that they intend to refer to the person. But do we say this because it was the priest’s intention to have this meaning conferred on the name? If we contrast case 1 with the others, this looks unlikely.

In case 2, after the ceremony, if the name ‘Aristotle’ is used in conversation by the members of the congregation, presumably it will successfully pick out the baptismal frock. The members of the congregation are in agreement over the meaning of the name—even though their shared beliefs conflict with the intentions of the baptizer. In case 3, if the members of the congregation use the name ‘Aristotle’, it will doubtless be used to pick out the baby grabbed from the priest’s arms before he collapses; subsequently, it will be used to call the boy to dinner; eventually, it will be used to address a grown man, just in case he survives for that long. In case 2, then, it seems that the name will come to pick out the person, even though the priest intended to baptize the baptismal frock. In case 4, it seems unlikely that it will be possible to use the name to refer to anything at all—if a speaker asks any other how it goes with Aristotle, they will not likely get an appropriate answer to their question.

Are we to say that in each of these cases, the name ‘Aristotle’ is conferred with the meaning intended by the baptizer? That the momentary intentions of the baptizer at the moment of baptism will override what may at least in cases 2 and 3 become a standard practice of naming for a linguistic community? If so, what do we mean by ‘meaning’ in these cases? What a certain speaker intended thousands of years ago that may conflict with every occasion of use of the name since then? For all we know, the name ‘Aristotle’ as we use it today could well have been uttered by a mad priest who intended to refer to a baptismal font two millennia past. Part of the problem here, I believe, is an ambiguity in the sense of ‘means’. I will turn to this next.

*Two Senses of ‘Means’*

To introduce an entirely different way of addressing the problem, I’d like to begin by pointing out a distinction between two ways that we use the word ‘means’, which have been distinguished in numerous places. Grice (1968) called the former ‘utterance meaning’, and the latter ‘timeless meaning’; Searle (1969) called the former ‘utterance
meaning’ and the latter ‘conventional meaning’; and Kripke (1977) called the former ‘speaker’s [meaning]’ and the latter ‘semantic [meaning]’\(^{39}\). While I don’t think that the criteria these writers have appealed to in order to distinguish the two are particularly successful\(^{40}\), they all have something in common. The essential aspect of the first sense of ‘meaning’ is tied to a speaker’s intentions, while the essential aspect of the second sense of ‘meaning’ is tied to a speaker’s ability to communicate with some audience using the term in question. The first sense of the term is captured in the phrase ‘what I meant to say was …’, while the second sense is captured by the phrase ‘in Italian, ‘pagliaccio’ means …’. These categories mark the difference between what speakers sometimes intend words to mean, and what those words can be used to communicate within some linguistic community. It is the difference between these two meanings of ‘meaning’ that is generating the conflicting intuitions in our debate. I’ll look at each in turn.

Now we can quickly see that a speaker’s beliefs are going to account sufficiently for meaning in the sense of intended meaning. All that is required for a speaker’s utterance to have an intended meaning is that the speaker intends it to. After all, it seems acceptable that a speaker would justify his having ‘meant’ such and such by any given utterance by saying ‘but what I intended to say was… ’. And so the reference-fixing descriptions discussed so far will account for the speaker’s intended meaning, which in the cases above corresponds to the priest’s intended meaning.

The second sense of ‘means’ is not reducible to speaker intentions. It is not sufficient for it to be true that in Italian ‘pagliaccio’ means clown that one Italian believes 'pagliaccio' means clown. If that was all that were required, then given the number of times Italian speakers are likely to have used words unintelligibly, or misspoken, or mislearned Italian words such that they used them in ways that nobody else could understand, it seems that we would have an enormous list of meanings for ‘pagliaccio’, or any other word of Italian. But this is certainly not what the authors of dictionaries

\(^{39}\) Specifically Kripke talks about ‘speaker’s reference’ and ‘semantic reference’, but the same considerations apply.

\(^{40}\) Grice’s distinction appeals to the claim that some terms in a language have a meaning that doesn’t change over time – but it seems doubtful that any terms have such permanent meanings as any dictionary will attest to; Searle’s distinction appeals to conventions to ground one aspect of meaning but we have seen that conventions are not necessary for terms to have meaning; and Kripke’s use of ‘semantic’ seems quite arbitrary: as we know ‘semantic’ simply means related to meaning – since the speaker meaning is also, obviously, related to meaning, contrasting it with ‘semantic’ meaning is not particularly helpful.
compile. The authors of dictionaries compile lists of the meanings that words can be used to communicate within a given linguistic community. And so, what is required for it to be true that in Italian ‘pagliaccio’ means *clown*, is that it is possible to communicate the meaning *clown* to other speakers of Italian by saying ‘pagliaccio’. Since this second sense of 'meaning' is essentially connected to communication, I will refer to it as the ‘communicative value’ of a term, adopting an expression used in linguistics. For the former sense of 'meaning', I would like to adopt the complementary phrase ‘cognitive value’.

*Shared Beliefs and Communicative Value*

Once we recognize that these two senses of the term ‘means’ play a familiar role in our use of the term ‘means’, I think it starts to become clear that a solution to the qua problem turns just as much on establishing the conditions for the determination of the second type of meaning as the first. If we consider in which cases the name seems to acquire some meaning in the second sense (that of ‘can be used to say’) and which cases it does not, we can see that there is a very clear common denominator. In cases 1, 2, and 3, there is no obvious reason why we would deny that the word ‘Aristotle’ has acquired a meaning. After all, for the members of the congregation in each of the cases 1, 2 and 3, there is a shared understanding of the meaning of the word. In case 4, however, it is not clear that the name ‘Aristotle’ could be used to communicate anything from one speaker to another. And in case 4, what is missing, is any shared understanding among the speakers. I would now like to propose that this shared understanding is just what underlies the second sense of ‘means’ that we considered a moment ago. What allows communication to take place is that speakers share beliefs about the meaning of the term. Indeed, it would seem that the extent to which speaker’s beliefs about the meaning of a term will be shared just is the extent to which speakers will be able to use that term to communicate with one another.

---

41 Van Patten defines the communicative value of an utterances as: ‘the relative contribution a form makes to the referential meaning of an utterance’ (Van Patten, 1996, 24), and by ‘referential meaning’ he means communicable meaning.
In order to solve the linguistic qua problem, then, we need to explore how pairs or groups of speakers manage to coordinate on a shared belief about linguistic meaning. We can see that the discussion at this point lines up with the discussion of radical interpretation set into place in Quine (1960). For any utterance, says Quine, it will be indi
cipherable to a radical translator – say a visitor from an unrelated linguistic community – what the utterance refers to. The possible references of a speaker's utterance of 'lo gavagai', announced in the obvious presence of a rabbit, will always be ambiguous between a range of possible referents. The meaning of 'lo gavagai' could be 'there's a rabbit', 'from here we have a good view of a rabbit', 'there's an undetached rabbit part', and so on. Equally, the reference of the utterance could be the object that is at a distance of twenty feet from the tip of the outstretched finger, or it could be the block of air immediately at the end of the finger.

To decide between these possible referents, Davidson (1992) proposed that the meanings of our utterances are established through a process of interaction between at least two speakers around a shared stimulus. He gives the following account of how we decipher a child's utterance 'table'. An adult watches a child's use of language, and recognizes a common stimulus accompanying each use. The adult sees a common pattern in the child’s utterances (the repeated word ‘table’), sees a table in the vicinity of each of the child’s utterances of ‘table’, and concludes that the child is recognizing tables as the reference of ‘table’:

The child finds tables similar; we find tables similar; and we find the child's responses in the presence of tables similar. It now makes sense for us to call the responses of the child responses to tables. Given these three patterns of response we can assign a location to the stimuli that elicits the child's responses. The relevant stimuli are the objects or events we naturally find similar (tables) which are correlated with responses of the child we find similar. It is a form of triangulation: one line goes from the child in the direction of the table, one line goes from us in the direction of the table, and the third line goes between us and the child. Where the lines from child to table and us to table converge, 'the' stimulus is located. Given our view of child and world, we can pick out 'the' cause of the child's responses. It is the common cause of our response and the child's response (Davidson, 1992, 119)

What is interesting about this approach is that it assigns meaning to the child’s utterances on the basis, partly, of the observer’s understanding. This fits with the model of
communicative value that I proposed above. The kind of 'meaning' that underwrites the utility of a language for communication, which is the kind whose values are compiled by the authors of dictionaries, is indeed dependent on facts about both the child and the adult’s beliefs. Once the adult and the child share a belief about the meaning of the utterance ‘table’ in each other’s mouths, and only then, can the pair use the term to refer to tables in conversation. The communicative values of their terms therefore emerge from the interaction of the pair, independently of the beliefs of either speaker.

Interestingly, recent psycholinguistics has begun to focus centrally on the role played by the shared environment in the acquisition of language, or what Davidson called the ‘triangulation’ process. Herbert Clark (1994) for example argues that it is only in the possession of a framework of mutual knowledge about one another that a pair of speakers can establish a communicative system. This mutual knowledge is established on the basis of what Clark calls the ‘co-presence’ of elements of the shared environment of the speaker and hearer. Tomasello (2003) has developed the notion of physical co-presence into what he calls ‘joint attentional frames’. A joint attentional frame is where the elements that are ‘co-present’ to both speaker and hearer are not only physical elements of the environment, but intentional elements. To introduce the notion of a joint attentional frame, Tomasello contrasts the following two cases. First, he considers a case where a monolingual English audience is addressed on a railway platform by a Hungarian speaker who presents the English audience with a string of Hungarian sentences. The audience, naturally, cannot understand a single thing that has been said – in our terminology we could say that the utterances have no communicative value for the audience whatsoever. This scenario is now contrasted with a case where a monolingual English speaker interacts with a Hungarian speaker at a ticket booth. In this case, the general field of likely references of the utterances of the Hungarian ticket master will be narrowed down by the joint activity in which the buyer and seller at the ticket booth are engaged in:

In this situation it is possible that the visitor may come to comprehend the communicative intentions expressed in some Hungarian words and phrases because the two interactants share an understanding of each other’s interactive goals in terms of gaining information about train schedules, obtaining a ticket, exchanging money, and so forth – goals expressed directly through the execution of meaningful and already understood actions such as the exchanging of ticket and money (Tomasello, 2003, 24)
Notice that it is not only the physical co-presence of the tickets and the money and the booth that allow the hearer to interpret the Hungarian speaker, but it is the ‘co-presence’ of the intentions of the speaker and hearer. Both speaker and hearer know that each one is at the booth in order to engage in a process of either buying or selling tickets. So the pair have mutual knowledge of each other’s intentions. This mutual knowledge allow each speaker to narrow in on the likely communicative intentions of the other in this context.

And it is a joint attentional frame, Tomasello proposes, that is the key ingredient in allowing infants to learn their very first words from their parents:

For example, suppose a child is on the floor playing with a toy, but also is perceiving many other things in the room. An adult enters the room and joins the child in playing with the toy. The joint attentional frame is those objects and activities that the child and the adult know are part of the attentional focus of both of them (ibid., 22)

When each speaker in a pair is attending to the same object, and knows of the other that she is attending to the object, the pair have mutual knowledge about the co-presence of that object to the other. But in addition to the mutual knowledge of the co-presence of objects, the pair acquire mutual knowledge of each others’ intentions by participating in some shared activity:

The joint attentional frames are defined intentionally, that is, they gain their identity and coherence from the child’s and the adult’s understandings of “what we are doing”, in terms of the goal-directed activities in which we are engaged. In one case we are playing with a toy, which means that certain objects and activities are part of what we are doing, and in another case we are changing a diaper, which brings into existence, from the point of view of our joint attention, a whole different set of objects and activities (ibid, 22)

Within this joint attentional frame, mutual knowledge is established between the parent and child of certain intentional states that both are in. Given true beliefs on the part of each partipant in the others’ intentions, he or she is in a position to interpret the others’ utterances (to begin with, the parent’s) as being likely relevant to those intentions. What this framework allows the child and the adult to do is to build up a stock of mutual beliefs that they can use to develop a system of communication. So we can see that a basic
process of collaboration underlies the initial acquisition of language, and fits very neatly with our current proposal for communicative value.\textsuperscript{42}

4. Shared Beliefs and the Qua Puzzle

If we return now to the \textit{qua} problem bearing this dimension of meaning – communicative value – in mind, we notice that the problem takes on quite a different form. We can remember that the second question involved in the qua puzzle was how a term is conferred with a meaning for the linguistic community to which it is introduce – not just in the mind of the baptizer. To return to the case of Aristotle, the priest, we imagine, announces ‘Aristotle’ with his hand on the baby’s head. Let us suppose the reference-fixing description in the mind of the priest secures the reference of his intended meaning as Aristotle the person. Given the physical co-presence of the baby, shared background knowledge of the function of a baptismal ceremony, etc. the audience (B) may, let’s suppose, almost perfectly predict the reference-fixing description and governing laws determining the priest’s intended meaning. Now, the intersection of their beliefs about the meaning of the priest’s utterance, since their beliefs are identical, includes the belief that the name picks out Aristotle the person. In this case, A and B will be able to pick out for one another Aristotle the person, using the name ‘Aristotle’.

But we should expect that often a hearer’s beliefs about the reference-fixing description for a speaker’s utterance will not match up perfectly with the reference-fixing description the speaker is employing. The beliefs of the speaker and his audience could

\textsuperscript{42} We could also approach this question from the framework of ‘pragmatic presupposition’ (Stalnaker, 1973). Stalnaker points out that in order to make sense of the utterances of another we often have to make assumptions about their beliefs, or about their environment. When an interlocutor says ‘there’s one now!’, for example, in order to interpret their utterance we have to presuppose (or ‘suppose’, given the utterance), that there is at least one object within the salient visual field of our interlocutor. Given this supposition, we can proceed to decide which object from among those we will expect to see our interlocutor is referring to. Or if someone says ‘Steve’s car is pretty fancy’, an interlocutor will have to presuppose that Steve has a car in order to make sense of the claim, in the same way. It is probably important to point out that the current proposal does not reduce to the pragmatic presupposition proposal, however. Pragmatic presupposition is employed to aid communication of intended meaning. On the current proposal, shared beliefs are constitutive of the communicative aspect of meaning. Recognizing that we often identify this aspect of language – what is constrained by the shared beliefs of interlocutors – as being constitutive of meaning, and not speakers’ intended meanings, is necessary to answer the questions that I am raising here.
fail to pick out a kind or individual with identical metaphysical boundaries, in spite of having overlapping elements. For example, if the governing laws that one speaker takes to determine the constitution of the object of his reference are slightly different to the laws the audience expect determine the reference, then a significant aspect of their beliefs will not be in coordination. In that case, according to the shared beliefs proposal for communicative value, the communicative value will pick out whatever entity or class satisfies the intersection of the regions picked out by the two distinct reference-fixing descriptions. For example, we can imagine that for A and B, each share beliefs about the reference of the term ‘Aristotle’ that it refers to a baby they have just seen baptized, but B believes that it refers to the baby only, while A believes that it picks out the person of whom the baby is an initial stage – i.e. B’s governing laws specify that the reference cannot transform into an adult, while A’s governing laws do not specify this constraint. In this case, the name ‘Aristotle’ will acquire the communicative value that allows it to pick out the baby, but not the person. We can see this if we consider the extent to which A and B will be able to communicate using the name. Once the baby has grown into an adult, if A asks B how it goes with Aristotle, B will presumably wonder how A could think that ‘Aristotle’ was still around. But if the term is uttered by A with the intention of drawing B’s attention to Aristotle while Aristotle is still a baby, reference will succeed43.

This model can be particularly clearly illustrated for the case of color terms. In figure 1, the predicted communicative value for the term ‘mauve’ for two speakers A and B is identified as the intersection of the sets A and B:

---

43 This might remind the reader of Burge’s prediction that a shopper at a butcher who orders brisket every week believing that brisket means ‘beef’ will in fact retrieve brisket just in case the butcher believes the term means ‘brisket’ (Burge, 1979, 80). Since brisket falls into the intersection of the sets ‘beef’ and ‘brisket’, the communicative value of the term ‘brisket’ for this particular speaker/hearer pair will be simply ‘brisket’.
If speaker A believes the term applies to colors in the range covered in set A, while B believes the term applies to the range in set B, the communicative value of the term will pick out the intersection of those two sets. To see why, consider a conversation between A and B around a box of variously colored pink and purplish crayons. A says to B ‘pick out the mauve crayons’. Clearly, B will pick out all the crayons that fall into the color range in set ‘B’. But A will insist that he only asked B to pick out the crayons in set ‘A.’ What we can see is that A was able to use the term ‘mauve’ to communicate with B only about the crayons picked out by the intersection of their sets of beliefs. B believed that A’s utterance ‘mauve’ referred to the colors in ‘B’ but not in ‘A’; but A’s utterance didn’t, because A didn’t intend it to. On the other hand, A intended his utterance to communicate to B a message about the range of colors in ‘A’, but not ‘B’; but A’s utterance didn’t, because B didn’t believe that it did.

The meaning A’s utterance had for the pair, then, is fixed by the intersection of their beliefs about that meaning. Similarly, the meaning conferred upon a term in an act of initial baptism for the linguistic community to which it is introduced will be fixed not by the intentions of the baptizer alone, but by the extent to which the community acquire shared beliefs about the meaning of that term. Qua what an object will be baptized, then, is determined by extent to which the governing-law reinforced reference-fixing
description in the mind of the baptizer, and his audience’s beliefs about what that reference-fixing description might be, intersect.

Conclusion

In this chapter I explored the inadequacy of an individualist model of meaning to deal with one crucial aspect of language – the introduction of new terms to a language. I argued that while facts about speakers taken in isolation are not sufficient to confer a determinate linguistic meaning in the sense of ‘communicative value’ on those speakers’ utterances, facts about speakers taken in pairs or group can support such attributions. In the next chapter I extend this analysis to linguistic practice more generally, showing that the two dominant accounts of the ground of linguistic meaning – intentionalism and conventionalism – are inadequate to account for the constraints that we find on linguistic meaning, and the fact that our use of language can be assessed for normativity, or, can be assessed as being either correct or erroneous.
Linguistic Norms – From Humpty Dumpty to Hunky Dory

“I could say oo-ee-oo-ee-oo-oo-oo, and everybody would know exactly what I was talking about”

Paul Simon, *Diamonds on the Soles of her Shoes*

In the last chapter I explored the deficiencies of individualist models of linguistic meaning for handling the problems surrounding the introduction of new terms to a language. I proposed that a way to rein in the indeterminacy of meaning that threatens the introduction of new terms is to appeal to the constraint of shared beliefs: an ‘initial baptizer’ can only introduce a new term to the extent that his audience is disposed to share beliefs with him about the meanings of his utterances. In this chapter, I develop the argument for grounding linguistic meaning in shared beliefs by exploring the debate between the two dominant accounts of linguistic meaning. These are, on the one hand, the view that speakers’ intentions are sufficient to ground the meaning of their utterances (e.g. Grice, 1957; Donnellan, 1967; Sidelle, 1992; Predelli, 1996, 2002; Akerman, 2009), and on the other hand the view that linguistic meaning can only be set in place by social convention (e.g. Strawson, 1964; Searle, 1967; Lewis, 1969; Schiffer, 1970; Corazza et al., 2001; Romdenh-Romluc, 2002; Gorvette, 2005). I explore and develop objections facing these views and argue that the views are untenable as they stand. I then build on the shared-beliefs model introduced in the last chapter to provide a model of linguistic meaning that avoids the problems facing both of these positions.

1. Intentionalism and its Woes
An early version of intentionalism and a Humpty-Dumpty worry are presented in Grice’s famous (1957) account of linguistic meaning and the reply given by Searle (1967). Grice proposed that the meaning of an utterance $U$ is $P$ just in case: the speaker of $U$ intends that his audience come to believe $P$ on the basis of hearing $U$, and his audience does in fact come to believe $P$ in this way.\(^{45}\) Searle presents a problem for this proposal as follows. Suppose I am an American soldier lost in the Italian Alps in WWII and I encounter Italian Axis troops. I think of a clever plan to get them to believe that I am a German Axis soldier – I will utter the only German sentence I know, and guessing that they will not understand German but recognize it as German when spoken, they will come to believe that I am a German soldier. So I say ‘Kennst du das land wo die Zitrönen bluhen’ and intend that they will come to believe that I am German by hearing me speak and recognizing my intention. My plan works, and I survive the encounter. But the same cannot be said for the intuitive appeal of Grice’s account: since I intend the Italians to come to believe that I am a German soldier on the basis of recognizing my intention that they would believe this as a result of hearing me speak, and I succeed, this example satisfies Grice’s conditions. Nevertheless, we have a strong intuition in this case that my utterance did not mean I am a German soldier\(^ {46}\) – rather it meant do you know the land where the lemon trees bloom. And so here we have a problem for the claim that my utterances get their meaning just on the basis of my intentions, however complex. If it were true, there would be nothing stopping me from saying ‘Kennst du das land wo die Zitrönen bluhen’ and meaning I am a German soldier, or we’re off to see the Wizard, or anything else for that matter, just as long as I intended to. This was the first hint of the Humpty-Dumpty problem. The next came in response to another intentionalist proposal given by Donnellan (1967).

Donnellan noticed that it is sometimes possible to refer to something using a description that is in fact false of the reference. His example was of a speaker who points

\(^{45}\) Grice’s claim is slightly more complex – the audience must not only come to believe $P$ after hearing the speaker’s utterance, but they must come to believe $P$ on the basis of recognizing that the speaker intend them to come to believe $P$ on the basis of hearing his utterance. I have simplified the model here for ease of exposition – the dialectic I am exploring is not affected by this simplification. For extensive discussion of this hypothesis and possible alternatives and variations see Strawson (1964); Searle (1967, 1969); Schiffer (1970).

\(^{46}\) Throughout I will use italics to denote utterance meanings and single inverted commas to denote utterance tokens.
to Jones, the man before him in the witness box, and says ‘the murderer of Smith is insane!’, believing Jones to be both insane and the murderer of Smith. While Jones is in fact not the murderer of Smith, our intuitions tell us that the speaker has referred to Jones nonetheless, and said of him that he is insane. But it is clear that it is not a relation of satisfaction between Jones and the description ‘the murderer of Smith’ that fixes Jones as the reference of the speaker’s utterance – since Jones is not the murderer of Smith. So what does? Donnellan suggests that it is the speaker’s intention to refer to Jones that determines the reference as Jones. In response to this proposal, Alfred MacKay (1968) raised the worry embodied in Lewis Carroll’s character Humpty Dumpty. Humpty is confident that his utterances mean whatever he wants them to mean – “neither more nor less” (Carroll, 1866, 119) – which he uses to justify his bizarre uses of language to Alice. MacKay worries that if our intentions are sufficient to fix the meaning of our utterances, then the deeply counter-intuitive implications of Humpty’s view would be true: a speaker could mean anything he pleases by any utterance. If a speaker’s intentions are sufficient to determine the meaning of ‘the murderer of Smith’ as Jones, then what is there to stop the speaker from intending his utterance to have any meaning that he cares, including the murderer of Jimmy Hoffa or the Twelve Days of Christmas? Now perhaps we might think ‘well, it’s certainly counter-intuitive, but perhaps our intuitions are simply wrong – perhaps I can mean whatever I want with any utterance’. It is true after all that so far all that the Humpty Dumpty argument amounts to is an illustration that some counter-intuitive implications arise for intentionalism. Shortly, however, I will articulate a stronger version of the Humpty Dumpty worry that goes beyond mere intuitions, and shows that a Humpty Dumpt theory of meaning divorces meaning from anything that can reasonably thought of as language. But first I will turn to the most recent incarnation of intentionalism.

In his (1992), Sidelle articulates a very nice puzzle for one straightforward account of the semantics of the term ‘now’. When we leave messages on our answering machines, we often say ‘I’m not here right now, but if you leave a message I’ll get back to you’. If as Kaplan (1978) proposes, the semantic content of ‘now’ always picks out the

---

47 Part of the motivation behind Donnellan’s paper was to provide counter-examples to Russell’s view that the content of a referring expression is always a disguised definite description that is true of the reference (Russell 1905, 1912).
time of utterance, then the proposition expressed by the message on the answering machine is patently false, even impossible, amounting to *I’m not here at the time of my uttering this sentence, but if you leave a message [etc.]*. For all that, callers do not balk in disbelief when hearing such messages, so it seems that the utterance has a coherent meaning after all. Sidelle suggests that in these cases, the speaker’s intention to refer not to the time of utterance but to the time when the message is heard alters the semantic content of the term ‘now’ to *the time of hearing* – an intentionalist account of the meaning of ‘now’. Predelli (1998) identifies another example of such a puzzle. If a professor knows that his colleague is away for the day and wants to warn students of his colleague’s absence, he might write a post-it note with the message ‘I’m not in today’ and stick it to his colleague’s door. In this case visitors to the office will all understand the inscription ‘I’ on the post-it note to refer to the usual occupant of the office, and not the writer of the note. And so here we have a counter-example to a straight-forward account of the meaning of the term ‘I’, which on Kaplan’s model always picks out the producer of the utterance. Predelli suggests, following Sidelle, that in these cases the writer’s intention to refer to the occupant of the office over-rides the standard meaning of the term. A more familiar example might be the inscription ‘wash me’, that we often see on the back of trucks that have had too much dirt thrown up on them from the street. In these cases, nobody takes the term ‘me’ to refer to the writer of the note, but to the truck. We can assume that according to Sidelle and Predelli, what fixes the truck as the reference of the inscription ‘me’ will be the writer’s intention. And so Sidelle and Predelli advocate intentionalist accounts of the semantics of the indexicals ‘I’ and ‘now’.

In response to the intentionalist proposals of Sidelle and Predelli, it has been argued (Romdenh-Romluc, 2002; Corazza et al., 2001; Gorvette, 2005) that the same Humpty Dumpty problems will arise. Gorvette puts the worry nicely:

> There would appear to be nothing to prevent the term ‘I’ from referring to one’s uncle, boss or the Pope if it is one’s intentions that determine the [content of these terms] (Gorvette 2005, 297)

Similarly, if our intentions are sufficient to fix the reference of our utterances of ‘now’ as in Sidelle’s proposal, we might have the same worry as Akerman:
It seems that John cannot, while standing on Trafalgar Square at 2 p.m. on the 15th of June 2007, use the sentence ‘I am here now’ to express the proposition that Bill was in Singapore at 5 p.m. on the 23rd of January 2006. But as long as there are no constraints on reference apart from the speaker’s intention, and no constraint on intention forming, intentionalism entails that John can express this proposition by using those words, since the intention is what determines the reference (Akerman 2009, 164).

In this way an intentionalist account of indexicals gives the same counter-intuitive results as the intentionalist accounts of definite descriptions given by Donnellan and that of language in general given by Grice.

**Intentionalist Replies**

So far the intentionalist has floated three responses to the Humpty-Dumpty charge. Donnellan’s response to the Humpty Dumpty charge is to argue that while it would be a counter-intuitive and unwanted result of a theory of linguistic meaning that a speaker could mean anything by any utterance, this will not happen on his model because a speaker’s intentions will be constrained by what she thinks is possible. Just as it is not possible to form the intention to fly by flapping one’s arms, Donnellan suggests, it is not possible to form the intention to mean it’s cold here by saying ‘it’s warm here’ (Wittgenstein, 1953, § 202). In each case, what stops us from forming the relevant intention are the limits on what we think possible. We don’t think it’s possible to fly by flapping our arms because we don’t believe our arm flapping will lift us into the air, and we don’t think it’s possible to say ‘it’s warm here’ and mean it’s cold here because we don’t think anyone would understand us if we tried. So according to Donnellan, while intentions are sufficient to determine meaning, the intentions that we can form are constrained in a way that blocks the Humpty Dumpty worries.

Predelli responds to the Humpty Dumpty charge with the suggestion that we need to distinguish that to which an utterance refers from that to which it can be ‘used to refer’. Predelli wants to say that while intentions are sufficient to fix what a speaker refers to, they are not sufficient to determine what a speaker can succeed in communicating to any given audience. So Predelli wants to retain a sense of ‘refers’ or ‘means’ for utterances
that may be incapable of communicating anything to any audience. In that case, Predelli urges, while the meaning or reference of my utterances will always be fixed by my intentions, they will not always determine what my utterances can be used to communicate to some audience. This distinction is possibly at the heart of the problem, but it needs to be explored in much more detail – so I will return to it later in the paper.

Finally, Akerman (2009) proposes something like a variation on Donnellan’s response. Akerman argues that a speaker’s intentions will be constrained by her habits in conforming to the linguistic practices of her community:

> It is in general hard for speakers to refer in the non-standard ways described in the Humpty Dumpty problem, since they are used to conforming to the external standards (at least to some extent), and in fact normally refer (roughly) in accordance with them (Akerman, 2009, 165)

The first thing that we can observe about these responses is that they do not rule out the possibility of utterances having Humpty Dumpty meanings – at most, they only reduce its likelihood. Donnellan, for example, argues that I won’t be able to form the relevant meaning-intention if I don’t think that it is possible to use a term in the deviant way – but gives us no reason why I couldn’t come to believe that the meaning in question is in fact possible. All that seems to be required to be capable of forming the intention to mean *it’s cold here* by saying ‘it’s warm here’, is that one believe that the expression ‘it’s warm here’ could be understood to mean *it’s cold here*. So it is certainly possible for utterances to have Humpty Dumpty meanings on Donnellan’s account – all it requires is that I have at least one false belief. Predelli, on the other hand, doesn’t attempt to eliminate the possibility of these meanings at all, but instead proposes to distinguish meaning from ‘being used to mean’. And while Akerman expects that the force of habits that I have picked up in my linguistic community will constrain me from ‘meaning’ in particularly deviant ways, he hasn’t provided any reason for us to think that we couldn't drop those habits, and admits that this is certainly not impossible.

So on the intentionalist account, the possibility of meaning anything you like by any term is not ruled out – at best it is simply made unlikely. I will now turn to a stronger

---

48 Akerman (2009, 165): “on the other hand, if a speaker believes that it is possible to refer in a certain deviant way, there is no reason to deny that she could form an intention to do so”
formulation of the Humpty Dumpty problem that goes beyond identifying merely counter-intuitive results for intentionalism, and shows that this remaining possibility of meaning in any way you like causes serious damage to intentionalism as a theory of linguistic meaning.

Even Humpty-Dumptier Worries

We have up to this point considered the Humpty Dumpty problem as the difficulty that some single utterance could be used with an indefinite number of different meanings. But as MacKay pointed out, the converse is also entailed by the Humpty-Dumpty implications of intentionalism – that it would be possible for a single meaning to be delivered by indefinitely many different utterances:

If one can refer to a book by using "the rock," then one can refer to a book by using any expression (MacKay 1968, 201)

Looking at the Humpty Dumpty problem from this perspective, we can see that it presents us with a much more slippery slope than before. At the top of the slope, we have cases where it is possible to refer for example to the murderer of Smith with the utterance ‘Jones’s murderer’, just by intending to. We can interchange expressions and by keeping our intentions fixed, the linguistic meaning of our utterances will remain the same. But now watch what happens. While we can exchange ordinary phrases of English on this model, it is also possible to distort phrases. I can replace ‘Jones's murderer’ with ‘Mones jurderer’ and mean the murderer of Smith, just as long as I believe that I can, as Donnellan has proposed. That nobody will understand this utterance is not relevant, because intentionalism does not take into consideration what other people can in fact understand – and only considers at most what I believe others can understand. Equally, I can say ‘Ones urderer’ and mean the murderer of Smith, just in case I believe I can. And so, armed with the belief that someone might understand me, I can say ‘ones urderer’, ‘nes erer’, all the way down to ‘…’, and mean the murderer of Smith. According to intentionalism, it is possible to sit in total silence and mean the murderer of Smith, just as long as your beliefs are sufficiently erroneous.
MacKay was worried that on Donnellan’s account ‘[meaning] is collapsed into the speaker’s intending to [mean], and the audience's (somehow) being able to get what he means’ (MacKay 1968, 201). But we can see that the situation is worse: intentionalism reduces meaning to simply thinking, and makes not just the choice of expression uttered irrelevant to meaning, but whether a subject speaks at all. Now we can see that the Humpty-Dumpty considerations don’t just produce counter-intuitive results – they illustrate that whatever sort of ‘meaning’ intentionalism provides sufficient conditions for, it is not linguistic meaning. Since the linguistic behavior that we actually engage in is irrelevant to the intentionalist account of meaning, intentionalism cannot be providing us with an account of linguistic meaning.

For all that there is surely something to the intuitions of the intentionalists, to which I will return in section 3. But for now, I want to look at what is considered to be the standard alternative to the intentionalist model – a conventions-based account of language.

2. Conventionalism and its Woes

To avoid the Humpty-Dumpty style problem presented by his devious trooper, Searle appeals to the role that conventions might play in determining linguistic meaning. He suggests that in order for an utterance to mean \( p \), it is not sufficient that the phrase be used with the complex intention provided by Grice: in addition, a convention to use the utterance in this way must exist in some linguistic community. Introducing this constraint would of course prevent my saying ‘it’s warm here’ and meaning it’s cold here, along with the possibility of meaning anything you care by sitting in silence, since there is no convention in existence to do these things. So the introduction of the conventionalist constraint solves the Humpty Dumpty problem.

Corazza et. al and Gorvette propose a similar alternative as a solution to the puzzles presented by Sidelle’s answering machine messages, and Predelli’s post-it notes. On their view, the term ‘now’ on an answering machine refers to the time of hearing the
message and not the time of recording because a social convention exists to use the term in this way, and the term ‘I’ on a post-it note picks out the usual occupant of the office for the same reason. Unfortunately, these authors don’t get into much detail on what such a convention might consist in, and so barring one recent objection given by Predelli, various important historical objections to conventionalist accounts have been overlooked. I’ll take a look at Predelli’s objection and its predecessors now.

Predelli’s Worry

In the recent debate, Predelli presents the following objection to the view that conventions govern uses of indexicals in post-it notes and answering machines:

> Are we really to believe that there are particular conventions regulating the use of indexicals in postcards, answering machines, recorded messages, written diaries, post-it notes attached to one’s door, daydreaming, re-enactments etc? [By this account] the conventional apparatus supposed to support our use would have to expand to unlikely proportions (Predelli 2002, 313)

It’s not clear why Predelli thinks these proportions are unlikely, but presumably he is appealing to the fact that we don’t go around striking up agreements to use indexicals on answering machines and post-it notes in certain ways and not others. If conventions could only be understood as explicit agreements, Predelli’s worry would certainly be legitimate. And indeed, this objection was raised in an even more serious form for conventions understood as explicit agreements by Quine (1936; 1969).

Quine’s Worry

Quine’s worry for a conventionalist model of linguistic meaning is that it would be not just unlikely, but impossible:

> When I was a child I pictured our language as settled and passed down by a board of syndics, seated in grave convention along a table in the style of Rembrandt. The picture remained for a while undisturbed by the question what language the syndics might have used in their deliberations, or by dread of vicious regress [...] An original founding of
language by overt convention is not merely unhistorical but unthinkable. What is
convention where there can be no [possibility] of convening? (Quine, 1969, xi)

The meanings of the terms in our language cannot, at least originally, have been fixed by
explicit agreement, because we will have had no language in which to strike up the
agreement. Given that conventions understood as explicit agreements cannot possibly
govern the meaning of all linguistic terms, the claim that conventions understood in this
way are necessary for linguistic meaning is clearly false.

However a way out of Predelli and Quine’s worries is made available to the
conventionalist by Lewis’ (1969) account of conventionalism. According to Lewis,
conventions can be put in place without explicit agreement. An early illustration of this
possibility is given in Hume’s rowers. Put several rowers in a boat, and quickly they will
synchronize their strokes. They do this because each knows that the most efficient way
for the rowers to get to their destination is to row in synchrony. And so they quickly
synchronize strokes, and settle on a particular pace. But all this is settled without any
discussion. To this extent, Hume proposed, we can call such a practice ‘a convention or
agreement betwixt us, though without the interposition of a promise’ (Hume 1740, III, ii,
2). Lewis supposes that linguistic conventions might be put in place in this way – as what
we might call tacit conventions. Suppose you come to my house and see me address my
dog as ‘Sparky’. Now that you know that I use the term ‘Sparky’ to refer to my dog, you
might say ‘Sparky’s a good boy!’ Now I will know that you know that I call my dog
Sparky, and you will know that I know that you know this. From this point on, we can
use the term ‘Sparky’ in conversation to refer to my dog. As such, a linguistic convention
has been established between us. But again, no explicit agreement has been made in order
to establish this convention. Since Lewis provides in this way an account of conventions
that dispense with the need for explicit agreement, Quine and Predelli’s objections to
conventionalism are quashed. Tacit conventions leave no traces, so a lack of evidence of
striking up agreements (à la Predelli) is not an indication that we have no tacit

49 Establishing the conditions for what is called mutual knowledge, where each of a pair of speakers knows
p, knows that the other knows p, and knows that the other knows that the first knows p. Mutual knowledge
is first identified as a mechanism that arises in solving coordination problems in the game theory studies of
Schelling (1960), is appealed to centrally as a mechanism in the conventionalist account of linguistic
meaning given in Lewis (1969) and Schiffer (1971), and in the psycholinguistics of Clark (1994) and
conventions; and since tacit conventions do not need a language in which to be struck up, the infinite regress problem (à la Quine) is avoided. But there are graver worries for conventionalism still.

Davidson’s Worry

The biggest problem for the conventionalist view is that it is quite clear that there are instances of linguistic ‘meaning’ to be found that don’t rely on any conventional precedent at all – tacit or otherwise. Davidson (1986) noticed such examples in the use of language made by Mrs. Malaprop, a character in Sheridan’s 1775 The Rivals who hopes that her daughter will make a ‘not altogether illegible’ candidate for marriage, lauds her guest’s manners as exemplifying the ‘pineapple’ of politeness, and touts her own skill in expressing herself in the ‘oracular’. Of course there is no precedent or convention to use these terms in these ways, but importantly, Mrs. Malaprop’s audience always understands perfectly well what she means. And it seems quite clear that it is on the basis of Mrs. Malaprop’s linguistic behavior that she is succeeding in doing this ‘meaning’ work. So, what Davidson provides us with is a counterexample to the claim that conventional precedent – either explicit or tacit – is necessary for an utterance to be meaningful. If this is true then we cannot appeal to conventions to block the Humpty Dumpty worry – since linguistic expressions that are clearly meaningful can be established without any convention. And of course there are far more common cases of unconventional uses of language that are perfectly meaningful and that don’t result from someone making crashing mistakes – neologisms (‘laptop’, ‘gluon’), puns (‘Parsley is gharsley’ (Nash, 1995)), and portmanteaux (‘smog’, ‘Brangelina’, ‘portmanteau’), for example, are all cases where words are used in a perfectly meaningful way apparently without conventional precedent.

However, it might be objected that in these cases the meaning of the new uses of terms relies on conventions governing aspects of their context. For example, in portmanteaux, the meaning of the word relies on the meaning of its parts – and the use of the terms out of which the portmanteau is constructed could plausibly have conventional precedent (‘smog’ is constructed out of ‘smoke' and 'fog', which both have conventional
uses, and so on). Similarly, Mrs. Malaprop may be relying on the conventional use of the ‘figures of speech’ she is trying (but failing) to employ, in order to get her meaning across: if she simply announced “Pineapple!”, for example, let’s say when her friend was trying to fill in a crossword puzzle for which the clue was ‘Apex or paragon’, she could hardly succeed in getting the meaning ‘Pinacle’ across.

However, we can illustrate even more clearly that convention is unnecessary for linguistic meaning, if we consider the following scenario. Suppose that two people acquire false beliefs about the standard meaning of some term, and they happen to acquire the same false belief. For example, let's say that I read the phrase 'Hunky Dory' on a packet of potato chips, and I acquire the belief that 'Hunky Dory' means great tasting – contrary to the intentions of the potato-chip manufacturer, and the beliefs of most English speakers, who expect the phrase to be understood as ‘just fine’, or something along those lines. Now, let's suppose that you overhear someone eating their dinner, and saying to the chef who has asked how the food is, 'Hunky Dory!' You too acquire the belief that 'Hunky Dory' means great tasting, on the basis of another misunderstanding. Now suppose that we run into each other, and engage in a conversation about food. I believe 'Hunky Dory' means great tasting, and so I say ‘Sprouts are Hunky Dory'. Clearly in this case there is absolutely no conventional precedent – either governing the use of 'Hunky Dory' to mean great tasting, or governing the context such that no meaning other than great tasting could be attributed to my utterance. I could have after all have been about to say that sprouts taste awful. Nevertheless, it's quite clear that you will understand me perfectly. And so in this case it would seem that my utterance will have the meaning great tasting – but without any conventional precedent whatsoever. Conventions are clearly not necessary for linguistic meaning.

For all that, the fact that conventional precedent does not impose a constraint on the ways in which language can be meaningfully used does not imply that we should therefore fall back to the intentionalist position. The Humpty Dumpty problem still remains – something is imposing a constraint on language even if it is not conventional precedent. After all, even those cases where language is used meaningfully without conventional precedent are themselves subject to constraints – malapropisms, puns, neologisms and portmanteaux cannot be coined simply by making noises with the right
intentions. ‘Bububu’ does not count as a neologism for if it is not raining I shall go for a walk just in case some speaker intends it to (Wittgenstein 1953, 18), and Mrs. Malaprop is not in such a position that she could say anything at all and pull off a malapropism: while the term ‘illegible’ seems to acquire the meaning ineligible in her mouth, she could not, I don’t suppose, have said ‘gumdrops’ and given it the meaning ineligible just by intending to. At some point long before a speaker has gone as far as sitting in silence, we will say that her utterances have become meaningless or nonsensical, no matter what her intentions might be. Something is constraining linguistic meaning, even if it is not convention. The question is what.

3. New Directions

It seems that we cannot allow that intentions are sufficient to put meaning in place, or we would find ourselves with a theory that makes no reference to linguistic behavior and hence does no work in explaining linguistic meaning. But it also seems that whatever is required to constrain linguistic meanings to avoid this upshot, it cannot be that of conventional precedent, since there are clear instances of linguistic meaning that do not appeal to or rely on convention at all. What is needed at this point is a new direction for exploration on this topic.

Gauker’s Accessibility Criteria

One candidate for such a new direction was recently proposed by Christopher Gauker (2008). Gauker suggests that the reason writers have appealed so frequently to intentions to fix meaning is that they were simply unable to think of alternatives. He then suggests, focusing on the case of demonstratives such as ‘this’ and ‘that’, that besides speaker intentions, there are any number of ‘accessibility criteria’ that we can appeal to in

50 On the basis of a rejection of intentionalism due to another very important argument from the informativity of utterances which I don’t have space to explore here.
order to determine the content of an utterance without the need for resorting to intentions. Gauker considers six criteria, including perceptual salience (say after a loud bang, someone says ‘that was loud’ – the salience of the bang picks it out as the reference of the term ‘that’); prior reference (if you have just referred to something in a sentence and then you say ‘that’s quite a point’, the reference will be the most recent in the discourse – this is known as ‘anaphoric reference’ in linguistics), and direction of pointing (if a speaker points while saying ‘that’, the reference is taken to be the object that intersects the line extending from the speaker’s finger).

Gauker is doubtless correct that these criteria play a role in our identification of the content of a demonstrative: in fact these criteria among others, and their role in reference fixing, have been carefully studied in psycholinguistics, as I discussed in the last chapter. But these criteria are not what we are looking for. These criteria provide for us the means of discovering what the reference of demonstratives in our linguistic community happen to be. But the appeals to intentions and conventions were designed to answer the question why the terms in our language have the meaning they do – not the question 'what do the words in our language mean'. We can after all perfectly well imagine a linguistic community where none of these criteria are relevant to the interpretation of a demonstrative – where, for example, the reference of the term ‘that’ is always fixed as the object nearest the speaker’s left foot. Similarly, while in our community the reference of a gesture of pointing is the first object that intersects the line that extends from the wrist to the tip of the extended finger, there is no logical reason why such a gesture might not pick out the object that intersects with the line going in the opposite direction – from finger to wrist (Wittgenstein 1953, §185). Or as Stanley Cavell notes, it’s not obvious why we wouldn’t simply attend to the tip of a speaker’s finger as he points, just as my cat does (Cavell 1969, 94). And indeed evidence from developmental psychology tells us that children learning how to use pointing gestures often confuse, for example, the significance of pointing to myself as meaning ‘I’, and pointing to you as meaning ‘you’ - they often reverse the two (Petitto, 1988). The point is that these demonstratives, just like all of the other terms in our language, could have had radically different meanings. And if they had different meanings, the ‘accessibility criteria’ would be correspondingly different. So we can see that Gauker’s proposal has
pushed the problem back one stage – we are still left with the question what fixes the meaning, and hence the accessibility criteria, for demonstratives or any other terms. And so we are still in need of an alternative to intentions and conventions.

A Shared Beliefs Model of Meaning

I would now like to propose that the constraint that we discovered in the last chapter will determine what meaning will be conferred upon a term in a process of initial baptism is the same constraint that is causing the conflict in the current debate between intentionalism and conventionalism. For any two speakers engaged in conversation, there is a range of beliefs that they share. This set of shared beliefs, I will argue, is what constrains linguistic meaning. In other words, I would like to propose that what is fixing linguistic meaning is not private intentions, or public conventions, but shared intentions, or shared beliefs. This model, I will propose, will avoid the difficulties of both the intentionalist and conventionalist view.

As we saw, the two senses of the term ‘meaning’ have been distinguished in numerous places and in detail. The essential aspect of the first sense of ‘meaning’ is tied to a speaker’s intentions, while the essential aspect of the second sense of ‘meaning’ is tied to a speaker’s ability to communicate with some audience using the term in question. The first sense of the term is capture in the phrase ‘what I meant to say was …’, while the second sense is captured by the phrase ‘in Italian, ‘pagliccio’ means …’. These categories mark the difference between what speakers sometimes intend words to mean, and what those words can be used to communicate within some linguistic community.

We can now see how these two senses of ‘means’ are motivating the apparently conflicting intuitions of the intentionalist and the conventionalist. The sense of ‘meaning’ in the phrase ‘what I meant by that was …’ indicates what a speaker intended to say. Now we can quickly see that intentionalism is going to account sufficiently for meaning in this sense. All that is required for a speaker’s utterance to have an intended meaning is that the speaker intends it to. After all, it seems acceptable that a speaker would justify his having ‘meant’ such and such by any given utterance by saying ‘but what I intended to
say was…’. And for this kind of meaning, the Humpty Dumpty considerations do not apply. I may find myself believing that it is possible to mean *you should get yourself a better job* by sitting in silence and raising one eyebrow. And if someone were to ask me ‘what did you mean by that eyebrow raise’, it is not obviously incoherent for me to reply ‘I meant *you should get yourself a better job*’. So there is clearly one sense of the term 'means' on which my intentions fully determine my meaning – the sense that is interchangeable with the term 'intends'. Equally, on this sense of 'means' it is perfectly reasonable to hold that you meant one thing while saying something entirely different – as Evans (1982, 130) considers, the auto mechanics student who states in class ‘a spark is produced electrically inside the carburettor' can be corrected by his teacher with the remark 'you mean the cylinder'. Here, the teacher attributes to the student the intended meaning *cylinder*, but the teacher is not thereby suggesting that the word 'carburettor' means *cylinder* just because the student thinks it does. And that takes us to our second sense of 'meaning'.

The second sense of ‘means’ is not reducible to speaker intentions. It is to this second sense of the word ‘means’ that the Humpty Dumpty intuitions apply, since there are constraints on what it is possible to communicate using any word. In the last chapter I considered how such sets of shared beliefs can be illustrated to fix the communicative value of color terms. Alternatively, consider a scenario where two speakers have distinct beliefs about the term ‘tree’. Speaker A believes ‘tree’ refers to adult trees and saplings, but speaker B believes the term refers to adult trees only. Now consider what happens when speaker A says to speaker B ‘look at the those pretty trees’, pointing to a row of poplars, at the end of which is a sapling. Speaker A will have intended to communicate to B that he thinks the set that includes the sapling is pretty; but of course he will only succeed in communicating that the set that includes the adult trees, but not the sapling, is pretty. Speaker intentions are clearly insufficient to ground the communicative value of a term, which can only be specified for at least pairs of speakers. And this, I would propose, is the constraint on meaning that motivates the Humpty-Dumpty intuitions. I can’t confer a communicative value on a term just by intending to, just as there is an aspect of the meaning of his utterances that Humpty-Dumpty cannot control, no matter how much he pays.
At the same time, this constraint on meaning does not rely on conventional precedent. It may be as a result of an explicit or tacit convention that a speaker and her audience acquire a set of shared beliefs about the meaning of some terms, but it needn’t be. A speaker’s audience could determine what the speaker believes her utterances mean from the context of the speaker’s conversation, without appealing to conventional precedent for the meanings of her utterances, and in this way arrive at shared beliefs about the meaning of a speaker’s utterances without relying on convention. This seems to be what is happening in the case of Mrs. Malaprop – her audience realizes that Mrs. Malaprop continually inter-substitutes similar sounding terms, and so whenever Mrs. Malaprop says something particularly bizarre, her audience interpret her utterance making allowance for that fact and taking the meaning of the saliently bizarre uses of language as having the meanings of similar sounding terms that would make sense out of the sentences she produces. In this case Mrs. Malaprop comes to the conversation with unusual (some would say false) beliefs about the meanings of some terms, and her audience update their beliefs in light of Mrs. Malaprop’s linguistic behavior such that the pair arrive at a set of shared beliefs, and Mrs. Malaprop’s utterances acquire a communicative value. Alternatively, shared beliefs can be acquired without even relying on interpretation from context – by simple coincidence of beliefs. This explains how the expression ‘Hunky Dory’ in the example considered above could come to means great tasting. Since both speaker and hearer have acquired the belief that ‘Hunky Dory’ means great tasting, they have a shared belief about the meaning of ‘Hunky Dory’ – and so it acquires the communicative value great tasting in their discourse, without any conventional precedent whatsoever. So we can see that shared beliefs are both necessary and sufficient for the setting in place of communicative value.

We can now consider how this model will treat our earlier cases. We can see that it will allow for indefinitely many unusual uses of terms, just as long as speakers are in a position to acquire shared beliefs about their meanings. The communicative value of ‘now’ on a telephone answering message is the time of hearing just in case the speaker

---

53 Davidson (1986, 101) calls the inferences involved in constructing such an interpretation a ‘passing theory’ of meaning on the part of the audience, and distinguishes this from the ‘prior theory’ – the beliefs that the audience have about the meanings of English words in advance of their re-evaluation under this encounter with Mrs. Malaprop.
believes that this is the meaning of the utterance in this context, and the hearer of the message shares this belief with the speaker. In this case, the acquisition of these shared beliefs may come about simply through a charitable interpretation on the part of the hearer, who assumes that the message leaver is not insane and guesses that the speaker expects the hearer to interpret ‘now’ to mean the time of hearing. Given this interpretation, the hearer acquires a shared belief with the speaker about the meaning of the speaker’s utterance of ‘now’, and a mini-dialect emerges between them. The communicative value of ‘me’ in the ‘wash me’ slogan on the back of a truck is the truck, just in case the writer believes this is the meaning the term when it is inscribed in this location, and the readers share this belief with the writer. Such a shared belief may be arrived at in a similar manner as the beliefs concerning the answering machine message are arrived at. While enormous flexibility in linguistic meaning becomes available on this model – much more than on a conventionalist model – it still imposes a constraint such that the Humpty Dumpty worries are blocked. Contrary to Donnellan’s view, it is not straightforwardly true that I can use term ‘I’ to mean the Pope ‘I’ just by believing that I can. If no audience believes the term ‘I’ refers to the Pope in the context in which I attempt to use the term with this meaning, the term will not have any communicative value at all. It may have a cognitive value for me – but that’s not all there is to language. And as our Humpty Dumpty intuitions tell us, as do the intuitions of the authors of dictionaries, what we are generally concerned to identify in linguistic meaning goes far beyond the cognitive value of terms for individual speakers.

Accordingly, we can define for any speaker S and hearer H the communicative value of a term t is P just in case:

\[ S \text{ believes } 't' \text{ means } P, \text{ and } H \text{ believes } (S \text{ believes } 't' \text{ means } P) \]

This shared-beliefs model of linguistic meaning does two things: it explains the constraints we find on linguistic meaning that make Humpty Dumpty’s theory false – which intentionalism cannot do – and it similarly allows for and explains the enormous flexibility we find in our linguistic use – which conventionalism cannot do. There are still intuitive cases, however, that the current model does not explain. It is intuitively possible
for pairs of speakers to be *both* wrong about the meaning of a term. I turn to this question in chapter 5. In the mean time, of course, this model also explains why Paul Simon can say ‘oo-ee-oo-ee-oo’, such that everybody knows what he’s talking about. In the previous line in his song, he tells the listeners that when he says ‘oo-ee-oo-ee-oo’, he’s talking about ‘diamonds on the soles of her shoes’. Given that his audience now shares a belief with him about the meaning of his subsequent uses of ‘oo-ee-oo-ee-oo’, he can say ‘oo-ee-oo-ee-oo’, and just as he predicted, everyone knows exactly what he’s talking about.

**Conclusion**

Here we have proposed in more detail how the communicative value of a term is established, and we have argued that the communicative value of a term provides the constraint on linguistic meaning that intentionalist models cannot provide, while allowing for a flexibility of interpretation that conventionalist models cannot accommodate. So far, then, we have argued that the space of shared beliefs between speakers can be recruited to reign in the indeterminacy of meaning that leads to the *qua* problem for initial baptism, and can also explain our intuitions that there are constraints on the ways in which we can meaningfully use language such that linguistic errors are possible. In the next chapter, I turn from the question of linguistic errors to the question of conceptual error – another area where language has been thought to play a role. I argue that while much of so-called ‘conceptual error’ can be handled without appealing to public languages, a certain class of errors can only be handled by appealing to a metalinguistic component of thought.
In the last chapter we considered how linguistic norms illustrate that there must be a speaker-independent dimension of linguistic meaning. I argued that the most plausible ground for that dimension of meaning is the set of beliefs speakers share. I now want to turn to debates surrounding the mysterious question of conceptual norms – how it seems to be possible to mistakenly apply your own concepts. The chapter consists of four sections. In the first I discuss examples of so-called conceptual error in the literature, and the challenge they present to dispositional accounts of conceptual content. In the second I discuss what I consider to be generally failed attempts to explain this phenomenon while keeping a dispositional account intact. In the third section I propose that we can, after all, explain many cases of conceptual error within a dispositional framework, by appealing to broader aspects of the rational framework of a thinker who applies these concepts. In section 4, I argue that although we can explain many instances of conceptual error within a dispositional account, there remain a class of errors that do not fit this account. To explain these we need to appeal to facts about the linguistic community to which the thinker belongs.

1. Varieties of Error
Three discussions of conceptual error that have independently generated a significant literature are Burge's (1979) discussion of linguistic deference, Kripke's (1982) discussion of Wittgenstein's 'rule-following problem', and Dretske's (1982) introduction of what has come to be known as the 'disjunction problem' for informational semantics. These three discussions boil down to a single argument against individualism about concepts, which I will call the 'infallibility problem'.

**Burge Cases**

Tyler Burge (1979) presents us with a puzzle for an individualist account of concepts with his famous discussion of the self-diagnosing Alf. Alf experiences aches in his thigh, and thinks to himself 'I must have arthritis in my thigh'. As we know, you cannot get arthritis in your thigh: arthritis is a disease that affects joints only, not all bones. So what we want to say, naturally, is that Alf has a false belief: that he has arthritis in his thigh. Now suppose, as we might also be inclined to, that it is the case that facts about Alf taken in isolation are sufficient to determine the reference of the term 'arthritis' as it appears in Alf's thoughts. If that's true, then it would appear that the content of the term as it appears in Alf's belief would be something like (at least) 'a disease that can affect all bones', since this is what Alf believes the term means. But now we have a problem. If the term 'arthritis' as it appears in Alf's beliefs has the content 'disease that can affect bones', then it's not, after all, the case that Alf has a belief about arthritis. Rather, he has a belief about some other possible condition, that we might call 'tharthritis' – which is stipulated as a disease that can affect the bones as well as joints. So rather than having false beliefs that involve the concept ARTHRITIS, it turns out that Alf has a true belief that involves the Alf-syncratic concept ARTHRITIS. In which case, we have failed to capture our original explanatory target: that Alf has a false belief about Arthritis.

**Kripke Cases**

Taking up Wittgenstein's (1953, §§140ff) discussion of rule-following, Kripke (1982) presents us with a puzzle similar to that of Burge. For any past utterance of a word,
Kripke challenges us to identify the facts about the speaker that determine the meaning of the word. Considering a past use of the term 'plus', Kripke considers various candidate facts about the individual speaking the word, and argues that there is nothing in these facts that could be used to decide whether the utterance meant plus, or a function very similar to plus but not identical:

Perhaps in the past I used ‘plus’ and ‘+’ to denote a function which I will call ‘quus’ and symbolize by ⊕. It is defined as:

\[
x \oplus y = \begin{cases} 
x + y & \text{if } x, y < 57 \\
5 & \text{otherwise}
\end{cases}
\]

Who is to say this is not the function I previously meant by ‘+’? (Kripke, 1982, 9)

Presumably, if the meaning of my utterances is determined by facts about me taken in isolation, then some of these facts will be able to answer this question. The ‘dispositionalist’ answer to the question is that what determines the meaning of ‘plus’ as I use it is simply whatever manner I am disposed to use it in. However, “The dispositionalist labors under [a] potent difficulty”, Kripke tells us: “Most of us have dispositions to make mistakes” (Kripke, 1982, 29). The problem that errors present for a dispositional account of meaning is, in a nutshell, that if my dispositions to use a term determine the meaning of the term, then it would seem those instances of use that we normally consider erroneous would count as veridical, since they will count among my dispositions.

For example, I know that I have a tendency to make the odd mistake when adding. Let’s say that out of every thousand calculations (being very optimistic), I have a tendency to give the wrong answer within a margin of error of 1. One way of describing my case would be as having a disposition to give the sum of two numbers 99.9% of the time, but to give the sum plus or minus one .1% of the time. But there’s the rub. If my dispositions to use the term ‘plus’ are to use it to give the sum of two numbers 99.9% of the time, but the sum plus or minus one .1% of the time, and if my dispositions determine the meaning of the term, then it seems that I don’t use the term ‘plus’ to mean plus. Rather, my dispositions tell us that I use it to mean something else – a function that gives
the sum of two numbers 99.9% of the time, and the sum plus or minus one .1% of the
time. Call it ruus.

Of course the dispositionalist can concede this point – the dispositionalist can
simply say ‘very well – that’s what I mean by ‘plus’ - ruus’. In adopting this strategy, the
dispositionalist is assuming that we cannot be wrong about the meaning of the term we
are using. This would be analogous to responding to Burge by saying that in fact Alf
means *arthritis* by ‘arthritis’ – Alf’s beliefs or dispositions or internal states however
you want to construe them determine the meaning of his utterances. And my internal
states determine what I mean by ‘plus’ – and if I mean *ruus*, so be it.

But once we realize that the problem graduates immediately to the level not just
of word meaning but of conceptual content (McGinn, 1984; Boghossian, 1989), we may
find this conclusion harder to swallow. If the content of my concepts is, just like the
meaning of my words, determined by my dispositions to apply those concepts in certain
ways, then it seems that not only will we be incapable of using words with the wrong
meanings, but we will be incapable of making any errors of judgments at all. If I am
inclined to judge of red things that they are red, but also of blue things in the dark, then
rather than attributing errors of judgment to me, the dispositionalist will have to attribute
to my concept *RED* the content *red or blue in the dark*. If I am disposed to judge of cows
that they are cows, but also of some Yaks, then rather than attributing errors of judgment
to me when I encounter a Yak and think 'there's a cow', the dispositionalist is forced
to attribute to my *COW* concept the content *cow or yak*, and suppose that it is never applied
in error.

*Dretske Cases*

Informational Semantics (IS) can be thought of as a sophisticated generalization of
Kripke (1980) and Putnam's (1975) proposals for causal models of reference. The idea
behind IS is that the content of a concept is whatever external stimulus causes the concept
to 'token'. The basic variable at work is the notion of information, which is defined as
follows:
An event A carries information about an event B just in case if B didn’t happen, A wouldn’t happen.

In other words, some event carries information about another event if the former wouldn’t come about without the latter. So as we can see, informational semantics is an essentially causal model of content. In this way, IS models our concepts as natural indicators of what they are concepts of. For example, if I have a HORSE concept, IS holds that this is so just in case this concept is triggered in my beliefs by encounters with horses. My HORSE concept tells me that horses are about when it goes off, because since my HORSE concept is whatever is caused by horses and horses alone, its triggering carries information about the presence of horses. The advantage of the IS model is that it provides an apparently clear route to naturalizing intentionality. Indicators of this type occur naturally, after all. A compass exhibits this ‘natural intentionality’ (Dretske, 1994, 21-28) because unhindered pieces of iron will naturally align themselves with the magnetic field of the earth, thus naturally indicating the direction of magnetic north. In this way the compass naturally ‘refers’ to magnetic north.

And now we can see how the problem of error appears once more. If our concepts are natural indicators, in the way that the needle on a compass is a natural indicator, then it makes no sense to say that our concepts could fail to be triggered by what belongs to their extension, or that they could be triggered by something that doesn't belong to their extension. On the IS model, the extension of a concept will be defined simply as whatever causes it to trigger. So, as a dispositionalist model, IS does not seem to be able to account for the fact that we sometimes make mistakes. This problem has been

---

54 This development of the causal theory was motivated particularly to solve the problem of ‘proximal causes’. To illustrate consider this case. Someone presses a doorbell, and the bell rings in the house. The ringing is supposed to carry information that the doorbell was pressed. On the simple causal theory, all that is required is that the pressing caused the ringing. But now have a problem – more than just the pressing of the doorbell caused the ringing. Additionally, an electric current ran down a wire. Not only that, but the current ran down one particular wire. Now suppose that there are several routes that the current ran down, and it runs down a randomly selected route each time. Whichever route the current runs down is part of the cause of the ringing. But does the ringing represent that the current ran down the particular route that it did run down? Clearly it does not. The IS model gets around this difficulty by stipulating that what the symbol (the ringing in this case) carries information about (or represents) is whatever had to happen in order for the ringing to take place. The doorbell had to have been pressed. But the current didn’t have to run down the particular route that it took – it could have run down any of the other routes equally effectively. So, the ringing carries information about the doorbell being pressed, but not about which route the current travelled.
characterized for IS as the 'disjunction problem' by Fodor (1990). Supposing that my cow concept is sometimes triggered by yaks, how is IS to avoid saying that rather than being in possession of a cow concept that is sometimes triggered erroneously, I in fact have a concept with the disjunctive content \textit{cow-or-yak}, which is always triggered veridically?

\textit{The Infallibility Argument Against Dispositions}

We can now see that Burge, Kripke, and Dretske cases all reduce to the same problem for dispositionalism, which we can capture in what I will call the infallibility argument:

\begin{enumerate}
  \item P1: If my dispositions determine the content of my concepts, then my dispositions will always be consistent with the content of my concepts
  \item P2: I can make conceptual mistakes, so some of my dispositions are not consistent with the content of my concepts
  \item C: Therefore, it cannot be the case that my dispositions determine the content of my concepts
\end{enumerate}

Now we might think that this argument only applies to causal theories – as may have appeared initially to be the case with the \textit{qua} problem considered in the last chapter – and so to be redundant for the current project where I have already adopted a descriptivist account of reference-fixing. However, a descriptivism grounded in phenomenal terms is ultimately also a descriptivism grounded in dispositions. We can describe regions of our environment in terms of their phenomenal properties, but we can’t describe the phenomenal properties in further terms (cf. Wittgenstein, 1953, §217: “Once I have exhausted the justifications I have reached bedrock, and my spade is turned. This is where I am inclined to say: “This is simply what I do’’”). Our ability to understand the contents of phenomenal terms must rely on a capacity to detect these properties, recall them, compare them when placed alongside one another, and so on. And so the content of these phenomenal terms will ultimately be fixed dispositionally.

I think that it is possible to establish a dispositional solution to the last of these cases – Dretske’s cases – but that Kripke and Burge cases demand that our understanding
of what a concept is goes beyond the dispositional states of individual thinkers. Next I will look at a series of attempts to solve the Dretske cases, and propose what I think is a more successful solution.

2. Selecting Among Dispositions

The key to solving the Dretske cases is to identify a principle that will allow us to select from among our dispositions, which are the 'veridical' and which are the 'erroneous' ones. I will discuss four attempts to do that here: the first two are ‘diachronic’ – appealing to facts about a thinker over the course of his existence (or even beyond, as we shall see); and the second two are ‘synchronic’ – appealing to facts that can be specified about a thinker at a single point in time.

2.1 Diachronic Proposal I: The Learning Phase Proposal

Dretske's proposes that if we take the objects that caused a concept to be tokened during its phase of initial acquisition as the 'veridical' triggers of the concept, we can distinguish among subsequent dispositions in a principled way. Suppose I acquire the concept ORANGE JUICE as a child as a result of exposure to orange juice and orange juice alone. At a certain age, on occasion I see a bottle of peach squash and think 'there's some orange juice'. Well, Dretske proposed, this is a mistake because peach squash doesn't belong to the set of things that caused my concept to trigger during the phase of initial acquisition of the concept.

The difficulty that this proposal runs into is that it seems to force us to beg the question about the content of the concept. The problem is that there is no obvious point at which we might decide when the learning phase ends, and the mastery phase begins (Fodor, 1990, 61-64). Suppose I spot some peach squash at age 6 and think 'orange juice'. Well, perhaps if we stipulate that the learning phase for my ORANGE JUICE concept ended at age 5 then we could say that this judgment was an erroneous one. But, if we stipulate instead that the learning phase ends at age 7, then peach squash falls into the class of
things that trigger my concept during the learning phase, giving the disjunctive concept ORANGE JUICE-OR-PEACH SQUASH. And so the judgment in question will be a veridical one. In other words, we can't decide when the learning phase ends without presupposing what the content of the concept is – and so the learning phase proposal cannot help us to decide between the veridical and erroneous triggers of the concept.

2.2 Diachronic Proposal II: The Teleological, or Well-Functioning Device Proposal

A subsequent proposal is the ‘well-functioning device’ proposal. Variants of this proposal appear in Matthen (1984), Millikan (1989), Papineau (1993), and Dretske (1994). Dretske’s explanation of the proposal, which I draw on here, is particularly clear. Suppose I am out mushroom hunting, and I eat a purple mushroom that makes me sick. As a result, I may acquire the habit of avoiding purple mushrooms. We can think of this habit as having a certain function – I acquired it for the sake of avoiding mushrooms that make me sick. The purpose or function of the habit, therefore, is to allow me to avoid mushrooms that make me sick. However, supposing that not every purple mushroom on the planet is poisonous, this habit is sometimes going to cause me to avoid non-poisonous mushrooms. In those cases, it seems, we should say something like ‘my poisonous mushroom indicating device was triggered in error’. Similarly, it could be the case that not all poisonous mushrooms are purple. In those cases, my poisonous mushroom indicating device is going to fail to go off – I will eat the mushroom and it will likely make me sick (Dretske, 1994, 475). However, the proposal faces serious challenges, which I am inclined to think are decisive.

First it is not clear that functional selection distinguishes between disjunctive and non-disjunctive conceptual contents (Fodor, 1990, 72-73). Natural selection does not select for features that would have the same function in every possible context – it selects for features that have the function of allowing a creature to survive in his current context. One version of the disjunction problem might be set up as follows. Suppose that a frog’s tongue snap reflex – which allows him to catch flies – is triggered by flies. We might say that the cognitive event that is triggered in the frog that induces the tongue snap is the frog's FLY concept. Now we can set up a disjunctive worry: that since the frog will also
snap at passing black dots – say a stray bee bee from a bee bee gun – we cannot tell whether the frog's concept is a FLY concept or a FLY-OR-BEEBEE concept. The teleological solution is to argue that since snapping at flies allowed the frog to survive, the frog's 'concept' must have the content FLY. But suppose we now put a frog who had acquired, we might stipulate, the disjunctive concept FLY-OR-BEEBEE, in the same environment. Will he also be selected for? Of course the answer is yes – if the things that fall under the class FLY-OR-BEEBEE are all flies in the frog’s environment, the frog will end up eating flies, not beebees. As Fodor puts it “Darwin cares how many flies you eat, but not what description you eat them under” (Fodor, 1990, 73). So we have no way of telling from the fact that the frog survived whether the concept the frog was deploying had the content FLY, or FLY-OR-BEEBEE, or BEEBEEs, since either one will be selected for in an environment where the passing little black dots are flies.

Another difficulty, identified by Braddon-Mitchell and Jackson (1997), is that a consequence of the teleological view is that speakers might not have any idea what their thoughts in fact refer to. As Jackson later argues:

The folk qua folk do not have opinions, let alone justified ones, about the theory of evolution; and even if they did, they do not qua folk have justified views about which structures are or are not selected for. It took serious research to show that the chin is not selected for (Jackson, 2006, 86)

Jackson and Braddon-Mitchell present the argument as a problem for attributing beliefs to others – if your evolutionary history determines what you’re thinking about, and I have no idea about your evolutionary history, then I have no hope of telling what you’re thinking about. We seem, however, to be able to ascribe beliefs with particular objects to others quite successfully, so the teleological story predicts a problematic outcome. Importantly, however, we can extend this objection into the case of the first person. Whatever about the beliefs of others, surely, we might suppose, we are in a position to know what we are thinking about ourselves. Now obviously there are cases where it is discovered that what we thought we were thinking about we turn out not to be (cases of error), but it’s not clear how far we can go with this.

55 For discussion see Papineau (1993) and Jackson (2006).
To return to the example of the frog once more, suppose that a frog is selected for his capacity to pick out flies. Now suppose that the frog is placed in a new environment in which there are no flies, and only beebees. The frog survives for some other reason, but develops all sorts of opinions about beebees (they are deceptive, not to be trusted, devious, etc.). Well, according to the teleological account, the frog in fact holds all these beliefs of flies (which it has never encountered, or within a generation may not have encountered for years) and not beebees. Similarly, it seems to be possible on the teleological account to discover that your horse thoughts are all about cows, or zebras, or plausibly about some animal that has been extinct for millenia. The same upshot follows for the learning phase theory: all of my beliefs which I think are true and about Orange Juice could turn out to be false beliefs about Peach Squash, just in case the concept of mine that is triggered when I encounter Orange Juice was originally acquired by encounters with Peach Squash. We could call this problem of ‘endless error’. Although the next two theories I will consider avoid determining the content of a concept in a possibly distant historical event, the problem of endless error will reappear for them.

2.3. Synchronic Proposal I: Tye's Optimality Conditions

Michael Tye has proposed a solution to the problem of error that may seem to avoid the difficulties of the views considered so far. Tye claims that a state $S$ in a subject $c$ represents a state of affairs $P$ if the following condition obtains:

$S$ represents $P = df$ If optimal conditions were to obtain, $S$ would be tokened in $c$ if and only if $P$ were the case (Tye, 2000, 136)

The optimal conditions clause is designed to accommodate the problem of error. Suppose I encounter a horse on a dark night, and think it's a cow. The horse triggers my cow concept, and we are presented with the problem of trying to decide whether I really do have a cow concept or whether the concept in question has the content cow or horse-in-the-night. Well, Tye proposes that we exclude this instance of triggering as being relevant.

---

to the content of the concept because it is taking place under sub-optimal circumstances. This seems reasonable.

The major difficulty with Tye's proposal is that it leaves open the question of what optimal circumstances are – and threatens to beg the question about the content of the concept in the way Dretske's learning-phase proposal did. It may be the case that, for some subject, we could decide that optimal circumstances are those that obtain in daylight when the subject is within 20 feet of the object of his judgment. If we decide this, then perhaps the subject's representational state $S$ will turn out to apply to cows and cows alone. But what if we are to decide that optimal conditions obtain when the subject is within 30 feet of the object of his judgment? It may turn out that, from 30 feet, both cows and yaks will trigger the concept under consideration. If we are to pick what counts as optimal conditions arbitrarily, then we will get different results about the content of the concept. On the other hand if we are to pick what counts as optimal conditions on the basis of ensuring a particular reference class for the concept, then we will be begging the question about the content of the concept. To decide on what optimal conditions are without begging the question, Tye (2000, 140) proposes to supplement the optimal conditions proposal with an appeal to a proposal of Fodor's – the ‘asymmetric dependence’ theory. However, the asymmetric dependence theory faces difficulties of its own, to which I now turn.

2.4. Synchronic Proposal II: Asymmetric Dependence

The asymmetric dependence theory is first proposed in Fodor (1987). Fodor’s idea is that we can distinguish content-determining dispositions from the rest, or veridical from erroneous dispositions on the basis of the asymmetric dependence of erroneous dispositions on veridical ones. Fodor claims that while I might sometimes mistake, say, a cow-in-the-night for a horse, my disposition to identify this cow-in-the-night as a horse depends on my also having a disposition to identify horses as horses. The reverse is not the case, however. My having a disposition to identify horses as horses does not depend on my having a disposition to identify cows-in-the-night as horses. Here’s Fodor on the
situation that holds, as he sees it, between erroneous and veridical tokenings of his HORSE concept:

The fact that cows cause one to say [or think] ‘horse’ depends on the fact that horses do; but the fact that horses cause one to say [or think] ‘horse’ does not depend on the fact that cows do (Fodor, 1987, p.108 – my square brackets)

Fodor's proposal is that the occasions of judging of cows that they are horses, or of garbage bags that they are dogs, depend on our ability to judge of horses that they are horses, and dogs that they are dogs. More precisely, Fodor specifies his proposal as a set of laws. "X" means X iff:

1. ‘Xs cause “X”s’ is a law
2. For all Ys not = Xs, if Ys qua Ys actually cause "X"s, then Y’s causing "X"s is asymmetrically dependent on X’s causing "X"s.57

This proposal faces a long list of objections, including that the model cannot handle twin earth cases (Dennett, 1987a, 1987b; Adams and Aizawa, 1994, 1997), that it cannot secure reference to individuals (Adams and Aizawa, 1994), that it cannot secure reference to non-existents, like Unicorns (Fodor, 1990; Seager, 1993). I have already discussed how we can secure reference to the actual world, and to individuals, and I am not going to discuss the problem of empty names here. Instead, I want to focus on and develop another widely observed objection and add something to it. It has been objected (Baker, 1989; Cummins, 1989) that Fodor's proposal does not do a good job of distinguishing between 'lookalikes'. Start off by supposing there are two species of butterfly nearby, monarchs and viceroys. These are mimics, and let’s suppose we are unable to tell them apart. It seems that if we are indeed unable to tell them apart, then it cannot be the case that our ability pick out one depends asymmetrically on our ability to pick out the other; if we can pick out either, we can pick out both. In such cases Fodor accepts the failure of

57 I am omitting from this formulation some extra conditions that Fodor includes to handle problem cases – specifically the claim that some non-Xs causes ‘X’ s, and that the dependence between 1) and 2) is synchronic, not diachronic – since I am not going to explore the problem cases that motivate these additions
my concept to distinguish between the two, and so accepts that in cases where we really can't tell members of two species apart, our concepts are disjunctive (Fodor, 1990, 102).

I now want to point out that Fodor’s response to this claim is in fact extremely problematic for his view: the kinds of thing that we mistake for, say, horses, just are lookalikes. Cows in the night are lookalike horses – which is exactly why we mistake them for horses. Garbage bags in the shadow of a wall just are lookalike skulking dogs, and again this is why we misjudge them. To get around this problem Fodor can argue that the 'covering laws' of his model apply to species in virtue of their essential properties, and not their accidental properties – thus distinguishing cows that only accidentally look like horses under certain circumstances, from horses, which essentially look like horses:

Being struck by lightning caused the death of the cow. The bolt that killed the cow was the fourth that Tuesday, so being struck by the fourth bolt on that Tuesday caused the death of the cow; “cause” is transparent to that sort of substitution. But though it's true that being struck by the fourth bolt on Tuesday killed the cow, the law that “covered” that causal transaction applies to cows and lightning bolts qua cows and lightning bolts; it was because it was a lightning bolt – and not because it was the fourth such bolt that Tuesday – that its hitting the cow caused the cow to die (Fodor, 1990, 102)

Fodor is making an appeal to the properties of the lightning that make it a good candidate for killing a cow to decide on qua what the lighting killed the cow. It isn't its property of being the fourth bolt on a tuesday that makes it a good cow killer. The last of things that come in fours on tuesdays don't all have in common the capacity to kill a cow. Its property of heating the column of air that it passes through to 20,000 degrees centigrade, on the other hand, will do nicely.

But now let's turn to the case of concepts. If we are to appeal to the qua of the causes of our concepts that makes them causes, then by far the most plausible answer for qua what the horse causes our concept to token, the result will not be horse at all – it will be qua horse lookalike. After all, it is looking horsey that all of the triggers of the concept have in common, and so if we select the qua by the procedure Fodor recommends – on the basis of identifying the property in virtue of which the trigger has the capacity to cause the concept to token – we will find being a horse lookalike is the obvious candidate: a property shared by both horses and some cows-in-the-night.
This brings us to the major chink in the armor of Fodor's theory: inevitably when Fodor talks about how we manage to get our concepts attached to their referents, he has to appeal to the superficial properties of those referents. Fodor talks about concepts attaching to their referents by a process that he calls 'locking on' (e.g. Fodor 1998a, 120-165). And in every account of locking on that he gives, the concepts lock on to their referents via the superficial properties of those referents. Of course this is the only plausible story Fodor could possibly tell – we can't access the hidden inner natures of the natural kinds in our environments just by standing next to them. Such a causal connection between a referent and a concept would be nothing short of magical. We learn about horses via their sensible properties. Eventually we develop a theory that tells us about the internal properties of horses too – but again, that's by opening up horses and looking inside. So concepts lock on to their referents via sensible properties.

If, however, we appeal to the sensible properties of things in our environment as the causal triggers of our concepts, we will all but have made the (highly plausible) retreat to the position that what the concept we have been discussing is caused to token by, in all cases, is the property of looking like a horse. This will imply that the concept being tokened is not a HORSE concept at all, but a LOOKS LIKE A HORSE concept. And of course, my LOOKS LIKE A HORSE concept is never triggered in error – since if I judge LOOKS LIKE A HORSE, then it is true that the object of my judgment looks like a horse to me (here, as in chapter 1, we are using ‘looks like a horse’ as a heuristic for a certain phenomenal Gestalt – the concept HORSE is not logically entailed by the concept LOOKS LIKE A HORSE). And in that case, we will be left with no explanation of semantic error.

Some have argued on behalf of Fodor’s theory in light of earlier articulations of this problem (Cummins, 1989) that there may be multiple mechanisms grounding the ‘X’s cause ‘X’’s’ law, so that even if some Ys caused ‘X’s because they looked like Xs, there may be other reasons why Xs would still cause Xs even if Ys didn’t (Manfredi and Summerfield, 1992). But once we see that it is not just looking like, but sounding like, smelling like, etc that will all succumb to this difficulty, the options on other mechanisms that might mediate the law attenuates fairly drastically. An appeal to alternate causes at this point will find itself faced with relying on ‘magical causes’ of the sort that might get
a horse to cause us to think ‘horse’ without seeing, hearing, or smelling one. Either way, I think there is an easier method of solving the problem of error, so I will turn to that next.

3. Categorical Versus Hypothetical Identification

My proposal in this section is that we can distinguish veridical and erroneous dispositions by appealing to the rational role that our concepts play in some and not other judgments. To begin with, we need to see that many of the cases of error that these authors are trying to explain are underdescribed. Taking a look at another older discussion of error will help to make this clear. Supposing George IV sees Walter Scott in the distance, and, unable to identify him, George IV thinks to himself 'Is that Scott?' (Russell, 1905, 485). To resist the absurd consequence of a purely referentialist account of the content of his thought as 'Is that that', or 'Is Scott Scott', Russell proposed that the two terms be thought of as expressing distinct descriptions, so that the content of George's wonder approximates to:

Is the man in the distance the man called 'Scott'?\(^{58}\)

Russell of course provides a descriptivist solution to the problem. But that is not what I am interested in at this point. Rather, I want to call attention to the characterization of George IV’s judgment that Russell has instinctually given, which shows us that there is a subtle difference between this case and the cases we have considered so far: in Fodor's cases, no room is made for the possibility that rather than judging definitively of cows-in-the-night that they are horses, what I am really doing is wondering to myself 'I wonder if that's a horse'.

---

\(^{58}\) This is not the description proposed by Russell to handle the case (see Ludlow (2003) for a comprehensive discussion), but it is similar enough to illustrate the point: suffice it to say that Russell thinks that if the descriptions that are cognitively equivalent to the referring terms in each instance are distinct, the problem can be handled.
The fact is that mostly when we are confronted with a blob-in-the-night that looks like it might be a horse, we don't judge simply 'horse!' Rather, we judge something along the lines of ‘looks like a horse, so it’s probably a horse’. What this shows is that there is more than one kind of thought that a concept can play a role in. In addition to what we might call ‘categorical identifications’, where we judge unequivocally of some stimulus that it falls under one of our concepts, there are also what we might call ‘hypothetical identifications’, where we put forward a probabilistic hypothesis about whether the object before us falls under one of our concepts. In many of these cases, it is not a HORSE concept that is directly triggered by the stimulus, but perhaps a LOOKS LIKE A HORSE concept. The HORSE concept is tokened subsequently as a result of the tokening of the LOOKS LIKE A HORSE concept. I think the key to solving the problem of error is in understanding how concepts can interact in this way. So first I’ll give an account of what I think the difference between ‘looks like’ concepts and ‘is’ concepts might be.

To explain the difference I would like to appeal to what psychologists call ‘superficial’ versus ‘essential’ concepts. For example, the prototype theory of concepts (Rosch and Mervis, 1975; Rosch 1973; Rosch 1975; Smith Shoben and Rips 1974; Hampton 1979) provides a carefully researched framework for talking about concepts that pick out classes of entities on the basis of their immediately observable properties. On the other hand, the ‘theory-theory’ model of concepts (Murphy and Medin 1985; Carey 1985; Keil 1989; Gopnik and Meltzoff, 1997) explores practices of classification that depend on ‘hidden’ features of the objects in our environment – such as their functional, genetic, or microstructural features.

It is generally accepted that we have both kinds of concepts: we understand perfectly well what it means for a thing to be ‘horsey-looking’, and the class of things that are horsey-looking for any individual thinker will be fixed just as the class of things the thinker judges to present the same immediately perceptible phenomenal Gestalt. ‘Essentialist’ concepts on the other hand have as their criterion typically some non-immediately discernible property of a class. For example, Keil (1989) explores how children from about the age of 8 will begin to classify things not just in terms of their immediately discernible properties but in terms of some ‘hidden property’ that they would have to investigate the object to identify. In one experiment, Keil shows children
drawings of a raccoon, and then an animation in which the raccoon is painted to look like a skunk, and has a sac of foul smelling fluid affixed to its hind (fig.1). Up to the age of about 8, children when asked whether the animal in the picture is now a skunk, will say ‘yes’; after that age, children begin to say ‘no’ – they now believe that what makes a skunk a skunk is not just looking like a skunk, but having been born with the features of a skunk perhaps, etc. What has changed is that the children have begun to classify objects in terms of hidden features of those objects, and not just their immediately perceptible properties.

![Figure 2: Raccoon painted to look like a skunk; reproduced with permission from Keil (1989) fig. 13.2](image)

It’s important to realize here that both the ‘superficial’ concept and the ‘essential’ concept are based on phenomenally discernible features of the object. The difference is that one set of phenomenally discernible features – looking skunky – is immediately discernible, while the other set – having some hidden property that could be discovered upon further investigation – is not immediately discernible. In the case above, the discernible but hidden property might be simply ‘can squirt me with a smelly spray’.59

What the general form of a judgment to the effect that something is, for example, a horse is likely to take, in that case, is not simply: HORSE! Rather, it’s tokening will be the result of 1) a visual presentation that looked horsey, which thereby directly triggered my LOOKS HORSEY concept; 2) an inductive generalization from previous encounters with horsey-looking objects that turned out upon further investigation to satisfy my

---

59 Obviously that’s not going to uniquely pick out skunks; but this doesn’t matter – for current purposes I’m only interested in showing how a concept whose extension is determined by our dispositions can be tokened in error – whether it picks out a natural kind or otherwise.
‘essentialist’ HORSE concept, to 3), the hypothetical identification of the object before me as belonging to the extension of my concept HORSE. Given such an inference from the appearance of a horsey-looking thing to the conclusion ‘probably HORSE’, I now go about my business representing my new environment as if it contains horses, because I have a reason to think that this is a good bet. And so I have a false hypothesis about my environment, that it contains horses. But the false hypothesis was not brought about on the basis of the false triggering of my HORSE concept in response to a cow. Rather, it was brought about on the basis of a hypothetical inference from the tokening of another concept historically associated with the first.

I would like to propose that this provides us with a way to handle cases of error without undermining the dispositional connection between the tokening of my concepts and causal encounters with their referents. Remember that the law the causal theorist is hoping to uphold is a law that holds between tokenings of the concept “X” and instances of that concept. Well, we can establish such a law, as long as we restrict it to instances of categorical identification as opposed to hypothetical. To justify this move, I need to expand on what I mean by the distinction between categorical and hypothetical identifications.

To do this I would like to appeal to something similar to Evans’ notion of a ‘canonical verification’ (Evans, 1982, 113). The idea is that for each of my concepts, there will be some possible epistemic scenario under which I could not doubt whether the concept applied. For some of my concepts, such as my ‘looks-like’ concepts, this will be every circumstance under which they are triggered. If something looks red to me, then I cannot doubt that the concept LOOKS RED applies to it. So for a concept like LOOKS RED, or LOOKS HORSEY, or LOOKS SKUNKY, every instance upon which they are tokened will be a canonical verification of the applicability of the concept, and hence a categorical identification. For looks-like concepts, every disposition counts.

On the other hand, if my concept is of an individual, there may be only one categorical identification that fixes its reference – perhaps on my first encounter with him – and all subsequent identifications may be hypothetical. To consider an example, I had a

---

61 For Evans this involves the verification for some claim ‘the a is F’, that d, a demonstrative, is identical with a, and that d is F (Evans, 1982, 113).
friend called Douglas when I was seven years old. We knew each other for a year or so, and then I left the school we were in. When I was about twenty, I met a man called Douglas. We both felt the other was familiar, and in conversation, we concluded that we must have been in the same school together. I had a good reason, in other words, to believe that this man was the person I knew when I was seven. The twist in the story, however, is that it turned out the man I met when I was twenty was not the boy I knew when I was seven. There were two boys called Douglas, and I subsequently re-encountered the one I had in fact known when I was seven. The point is that even though the encounter with the imposter Douglas – along with my judgment ‘DOUGLAS!’ – may appear to have all the hallmarks of a categorical identification, it in fact was not, since it was subsequently intelligible to me that he was not DOUGLAS after all. What makes the difference between a categorical and a hypothetical identification, in that case, is the possibility of doubt. In a canonical verification, my capacity for doubting whether my concept applies will be exhausted; in a hypothetical verification, it will not.

This provides us with a clear way of testing whether a judgment is a categorical identification or a hypothetical one. In chapter 1 I described how it was possible to test whether two referring expressions were cognitively equivalent by placing them in an identity claim and questioning whether the identity claim was conceivably false. We were able to show (after Frege) that the referring expressions ‘Jekyll’ and ‘the man who tips his hat in the morning’ were not cognitively equivalent. The sentence:

1) Jekyll might not be the man who tips his hat in the morning

is not incoherent, but when the description on the right was swapped with the name ‘Jekyll’:

2) Jekyll might not be Jekyll

the sentence was rendered incoherent, showing that the description and the name were not cognitively equivalent. We can now apply the same procedure to demonstrative identifications to distinguish between categorical and hypothetical identifications. What
we are looking for is a symptom of our capacity to doubt whether the concept applies to the object we are judging it applies to. If we can’t doubt it, then the judgment is categorical. If we can, it’s not. I argued above that an application of my LOOKS RED concept, or any other ‘looks-like’ concept, will be incorrigible. If that’s true, then a judgment of the form:

3) I think this looks like a horse, but it might not look like a horse

ought to be inconceivable, and I think it clearly is. In that case, ‘looks-like’ judgments will always count as a canonical verification of the applicability of the concept, and hence as a categorical identification of the extension of the concept. Concepts like ‘looks red to me’ will therefore have multiple categorical identifications.62

On the other hand, if an identification is to count as a hypothetical and not a categorical one, it will admit of the conceivability of error, which will be evident in the fact that a sentence that expresses the possibility of the misapplication of the concept in that context will be coherent, as the following sentence ought to sound:

4) I think that’s Douglas, but it might not be Douglas

Since 4) is perfectly coherent, it cannot be the case that my judgment ‘that’s Douglas’ is ‘categorical’, and so as confident as I may be of the truth of my judgment, it still counts as a hypothetical judgment. What this shows is that the judgment ‘I think that’s Douglas’ in this case is really a shorthand for ‘I think that’s probably Douglas’. Categorical identifications for terms like proper names will likely only occur on the initial occasion of reference-fixing of the name which I explored in chapter 1. And on the initial occasion of reference-fixing, of course, it will not be possible to doubt whether the concept applies to the object to which it has been descriptively fixed.63

62 Although this does not exclude the possibility of a looks-like concept entering into a hypothetical identification: if I am told there is can of Coke in a paper bag before me, I may well think ‘this will look red to me’. I could of course be wrong, if for example it’s a Cherry Coke
63 One possible worry for such a model, raised as an objection to asymmetric dependence theories, is that of so-called pathological cases, which could be constructed as an objection to the current model. Suppose a stimulus causes a concept to token without any sensible input – for example if an electrode were inserted
And so the following picture emerges. Our concepts play roles in two distinct kinds of identification judgments: categorical and hypothetical. It is possible for hypothetical judgments to involve the tokening of a concept not in conjunction with its actual referent – for example when I judge of a thing that triggers my LOOKS HORSEY concept that it might be a HORSE. We can tell the difference between categorical and hypothetical judgments on the basis of our capacity to doubt their contents. If we cannot doubt whether a concept of ours applies in a certain situation, then that situation counts as a canonical verification of the applicability of the concept, and is therefore a categorical identification. Such will be the case for LOOKS LIKE concepts in general. Proper names or concepts of individuals or ‘essential’ concepts, on the other hand, may have fewer or only one categorical identification. In some cases this will be the original moment of reference fixing, in others it will be based on the discernment of a ‘hidden’ property after some investigation that counts as the criterion for membership in the class. The toy SKUNK concept I discussed above, for example, will admit of multiple categorical identifications, since a canonical verification for this concept would simply amount to witnessing a creature spray from its behind.

The immediate reaction this proposal is likely to illicit is an objection of the form ‘but I can entertain the thought like ‘squirts from behind, but might not be a skunk’’. That is, it might be observed that we very often do not have a phenomenal criterion for determining membership in a class that we are absolutely confident determines membership. However, these cases are plausibly all cases of deference, which I discussed into my brain and stimulated some region that caused me to think HORSE (Baker, 1989; Seager, 1993; Adams & Aizawa, 1994). This would appear to be an instance of concept tokening falsely, but not on the basis of a hypothetical inference from the tokening of another concept. However, bearing the notion of a canonical verification in mind, the current model can handle these cases. In the case of incorrigible concepts like LOOKS RED, an electrode in the brain will at best produce an experience that looks red, thereby tokening my LOOKS RED concept. But of course such an experience will in fact look red to me, so the concept will not have been tokened in error. On the other hand if the electrode is to produce an hallucination of a horsey looking thing, that causes me to think HORSE, this will plausibly be via inference from the production of the hallucination, and so will also not count – in both of these cases it will be conceivable to me that I am hallucinating, and so the identification will not be categorical. A final construal of this challenge would be to suggest that the electrode could cause me think HORSE somehow ‘on its own’ – not in reference to any apparent perceptions, while also somehow depriving me of the capacity to suppose that I might be making a mistake; however, at this point we seem to be talking about rendering someone mad – and so I don’t think this is a reasonable counterexample, since I am inclined to think with Brandom (1996, 2000) and Sellars (1956) that a creature cannot be attributed a concept who cannot also play the game of ‘giving and asking for reasons’; see also Chalmers (2002b) who denies as being relevant counterexamples based on creatures who exist ‘not in the space of reasons’
in chapter 1: when I think ‘looks like a skunk, smells like a skunk, sounds like a skunk, but might not be a skunk’, the term ‘skunk’ here really means ‘the thing called ‘skunk’ in my community’.

Another problem that may appear to arise would be that of Endless Error that I identified for the diachronic proposals considered in section 2 – however, the current model deals with these cases very differently – the current model is synchronic, not diachronic. What counts as a canonical verification depends on my dispositional capacity to assess whether a scenario counts as canonical verification – it appeals to my current dispositional rational state. Consider the ORANGE JUICE/PEACH SQUASH case once more. If I acquired a concept J in conjunction with a presentation of peach squash as a child, and shortly after mistook Orange Juice for Peach Squash, and forever after that continued to do the same, simple historical instantiation condition theories will predict that I in fact have endless false beliefs about Peach Squash in spite of having not encountered Peach Squash for decades, which is deeply counterintuitive. The current model importantly avoids this upshot. The reason is that if I do not know that I originally encountered Peach Squash and not Orange Juice, then it is Orange Juice that will play a role in a canonical verification of a claim about J in my thoughts: what plays a role in a canonical verification is what I will accept as playing that role. Distant original triggers that are out of epistemic reach for me will not enter, for that reason, into a canonical verification of the concept’s applicability. We could see this happening through memory failure, for example. Suppose I met Douglas as a boy, and my DOUGLAS concept had its reference fixed descriptively to the boy. Now suppose I mistake another boy for DOUGLAS the next day (let’s say it’s the other Douglas). I become friends with this boy and stay in touch for twenty years. Do I have endless false beliefs about the second Douglas and little to no beliefs about the first? On the current theory this does not follow: as the original stimulus fades into history and out of my epistemic reach, a more recent encounter with the subsequent trigger will take the place of the original trigger as the canonical verification\(^\text{64}\). If what I will accept as a canonical verification of the applicability of DOUGLAS involves Douglas2, then Douglas2 is the reference of my concept.

\(^{64}\) This in fact I think qualifies as a cognitive analog of reference shift, a famous example of which is provided in Evans (1977) of the name ‘Madagascar’: having originally been the name for a part of the
The cases that would count as cases of genuine semantic error, then, are cases where my concept is categorically and not hypothetically applied. I think it is quite clear that in none of these cases will we have a failure of correspondence between the content of the concept and the stimulus that causes it to token, since the categorical identifications are stipulative of the content of the concept. The law that holds for conceptual content is therefore:

\[ S \text{ represents } P =_{df} S \text{ would be tokened categorically in } c \text{ if and only if } P \text{ were the case } \]

Identifying the ‘categorically’ clause allows us to exclude cases of the tokening of a concept as a hypothetical identification on the basis of the tokening of another concept.

This offers us a way of dealing with the general infallibility argument against dispositional theories of content. Premise 2 is false: it is not the case that our concepts are tokened by objects that do not belong to their extension. Rather, some of our concepts can enter into hypothetical judgments about our environment, given the tokening of other concepts and a history of association between the two. However, there is still a class of errors that this model does not account for. While we have refuted the general infallibility argument against dispositionalism, we can still articulate what we might call a special infallibility argument against dispositionalism – which is illustrated by the cases considered by Kripke and Burge.

4. Language and Public versus Private Concepts

---

African mainland, Marco Polo mistook the name as picking out the offshore island. The name clearly now refers to the island. The question is how that is the case given that the causal chain of naming originates with the baptism of a part of the mainland. Evans’ answer is that the reference of the name is whichever object is the causal source of most of the information that speakers file under their ‘Madagascar’ beliefs. I don’t particularly like this theory since it implies that the reference of a name could suddenly switch just as soon the range of my beliefs deriving from one object and not another tips over the 50% mark. This does not seem consistent with Evans’ overall approach, which is to always ground reference out in personal level rational states. The tipping of the balance of my beliefs over 50% from one source to another is a sub-personal event that I will have no awareness of, so this condition seems to violate Evans’ overall rationalist approach, and plausibly violate his Generality Constraint.
So far we have established that many cases of errors in judgment can be explained without violating the causal connection between a concept and its reference. In this way, we have shown that dispositionalism is in many ways much better off than is often supposed. The general infallibility argument against dispositionalism is unwarranted, once we realize that misrepresentation can arise entirely independently of semantic error. However, there are still cases of apparent semantic error that cannot be assessed along the lines that we have provided so far.

4.1 Getting Concepts Wrong

When Burge's self-diagnosing patient goes to the doctor, it turns out that not only his ‘feels like’ conception of arthritis is false, but his deep ‘arthritis’ concept is also wrong. Burge's patient believes that arthritis is a disease that can affect all bones, when in fact it can only affect the joints. If neither Alf's dispositions to classify arthritis according to superficial properties, nor his dispositions to classify it in terms of its ‘essential’ properties match up with what we consider to be arthritis, then why do we say Alf is thinking of arthritis at all? To have a false belief about arthritis, after all, Alf would have to be thinking of arthritis, not that arthritis.

Kripke's mathematician presents a similar puzzle. Kripke's mathematician is supposed to be applying a concept with the content plus, but his dispositions are consistent with his applying a concept with the content quus. Can we attribute to him to a deep concept of PLUS that he is inferring applies on the basis of it apparently applying? It isn’t clear how we might do this. I know how to add numbers, but I can’t say that I could provide you with any deeper criterion of what counts as ‘plus’ for me than simply to show you what I do when I am adding numbers.

These cases, it would seem, are essentially connected with our commitment to apply what we might call 'public concepts' in our thoughts, whose content is determined not by individual dispositions, but by shared beliefs among groups of speakers. And we go about this by tracking the linguistic behavior of other members of our community, and doing our best to conform with that behavior. Where we fail to conform, the community assesses our beliefs as erroneous, and where we succeed, as veridical. This will imply
that the content of thoughts like 'I believe I have arthritis' has a metalinguistic component. In the next chapter, I will take up the question of how the contents of 'public concepts' is determined, but for now we need to give a coherent account of the metalinguistic component of content.

4.2 Metalinguistic Analysis I

One metalinguistic approach to the cases of Alf and the deviant mathematician is to accept that Alf has his own idiosyncratic concept, but has false beliefs about the terms of English that can be used to express those concepts. On this approach, Alf's concept has the content \( \text{tharthritis} \), and Kripke's errant mathematician means \( \text{quus} \). We can call this the semantic infallibilist position, and it has been held by Block (1980), Fodor (1987), Crane (1991) and Segal (2000).

Crane, for example, holds that Alf says “I have arthritis in my thigh” because he believes that this is the right sentence to use to express a belief about tharthritis:

Alf has a true belief, \( I \text{ have tharthritis in my thigh} \), [and] a false belief to the effect that \( “I \text{ have arthritis in my thigh}” \text{ is the right sentence to express this belief} \), and thus makes a false statement, “I have arthritis in my thigh” (Crane, 1991, 298)

Crane argues that if we look at the psychological evidence, we will only find evidence that suggests that Alf has the concept \( \text{THARTHritis} \). If a psychologist were to examine Alf she would find that Alf applies the concept in question to conditions affecting bones as well as joints, she would find that Alf makes the inference from ‘I have condition X’ to the conclusion ‘it could be in my thigh’, and so on. (Cf. Block, 1980; Fodor, 1987, 40).

This analysis is far from satisfactory however: it makes nonsense out of the conversation Alf has with his doctor. As Tye puts it:

When the doctor tells the patient that he cannot have arthritis in his thigh, the patient, except in extraordinary circumstances, will not insist that the doctor is wrong and that he does have arthritis in his thigh. But on the meta-linguistic proposal, this is how he should respond (Tye, 2009, 66)
If Alf believed he had tharthritis, and not arthritis, then he ought to insist that his doctor has misunderstood him – he ought to say “oh, yes you can get arthritis in your thigh – by ‘arthritis’ I mean the condition I have in my thigh”. Obviously this would be an absurd reply, since it would make nonsense out of Alf’s bothering to have asked his doctor about arthritis in the first place. To make the point clearer still consider Alf asking his doctor ‘what’s arthritis?’ Suppose Alf believes that he is asking about a condition that can affect your thigh as well as your joints. On the semantic infallibilist analysis this would imply that Alf is asking about tharthritis. So when Alf asks ‘What’s Arthritis?’, and his doctor replies ‘it’s a condition that affects the joints alone’, Alf ought to reply, ‘no, it’s not, it’s a condition that affects the bones and the thighs’. Again, this is an absurd reply. The lesson is that Alf has not gone to his doctor to get a lesson in medical terminology – he has gone to seek advice about real medical conditions that he has heard of under names like ‘arthritis’, and that he is worried he has developed.

4.3 Metalinguistic Analysis II

A more successful version builds on the account of reference-fixing that we established in chapter 1. I argued there that in many cases the reference-fixing description that is cognitively equivalent to a referring term is simply ‘the reference of ‘t’’, where ‘t’ is a term I have heard of from someone else. This is the most straightforward analysis of Kripke and Burge’s cases.

The problem with Crane's proposal is that it suggests that what the speaker “really means” when he says 'I have arthritis' is that he has tharthritis. This leaves us with no explanation for why the speaker concedes to his doctor that he had a false belief about what the condition in his leg actually was. The meta-linguistic description on the other hand, places whatever it is that the word 'arthritis' refers to in the English language as the primary object so to speak, of the patient's judgment. In that case the patient is judging of the condition that is referred to by the term ‘arthritis’ that it has occurred in his thigh.

This second version of ‘meta-linguistic’ content can get us out of our puzzle. A thought like ‘Arthritis is a condition that can affect bones’ now turns out as:
the reference of 'arthritis' is a condition that can affect bones

Since there is no analytic or necessary connection between ‘the reference of ‘arthritis’’ and ‘condition that can affect bones’, someone who thinks this thought can concede to correction from his doctor, or anyone else, that it might be false. The thoughts of the deviant mathematician, similarly, can be captured by the claim that he is making judgments not about the results of his calculations that they conform with his own dispositions, but rather that he is making judgments about the results of his calculations that they conform with the function picked out by the reference of the term 'plus' in his community:

\[ n \text{ is the product of the numbers } o \text{ and } p \text{ under the function that is the reference of 'plus'} \]

Since it is perfectly possible that the reference of 'plus' in the language I speak is not the function that I am always disposed to compute when asked to do some plussing, this analysis allows us to accommodate the possibility of error while at the same time connecting the thoughts of the mathematician to the actual function plus.

What is important and interesting to notice at this point, however, is that we do not speak of Alf as failing to understand a word: we talk about Alf as failing to grasp a concept. This suggests that there is an ambiguity in our notion of ‘concept’ between contents that are fixed by the dispositional states of individuals, and contents that are somehow fixed by facts about a thinker plus her linguistic community. For example we speak of ‘failing to grasp the concept of Schadenfreude’, or ‘finally grasping the concept of duty’. It is a part of our ordinary discourse to speak of concepts as if they are the kinds of things that certain terms are names for, and that individuals can fail to understand, or partly understand (Cf. Burge, 1979, 136 “what I have called “partial understanding is common or even normal in the case of a large number of expressions in our vocabularies”). We also speak of concepts as individual dispositional states – when we say, perhaps, that my concept of happiness differs from your concept.
What all this suggests is that there is an ambiguity in the term ‘concept’ that parallels the ambiguity in the term ‘meaning’ that we explored in chapters two and three. When we talk of grasping the concept ‘arthritis’, we speak in a way that presupposes that an individual thinker could fail to grasp the concept. The analysis of such cases that we have proposed here is that the content that the speaker is failing to grasp is determined by the beliefs of her community, as determined by their linguistic use, or perhaps the opinion of experts in her community. We might rightly speak, in that case, of a distinction between what we could call ‘private’ and ‘public’ concepts. The content of private concepts will be fixed by the mechanisms we have explored in chapter 1, 2, and above, all of which can be specified by facts about thinkers taken in isolation. What fixes the content of ‘public concepts’, via the use of terms in a shared language, is a question that I turn to in the next chapter.

**Conclusion**

In this chapter I considered the classic argument against individualist accounts of concepts from error. I argued that the argument in its 'infallibility' form it does not succeed against individualism, since we can construct a dispositionalist account of content that allows for misrepresentation without violating the causal connection between concepts and their referents. I then considered that a weaker form of the argument from error persists as a difficulty for individualism – that some kinds of error still outrun the individualist account. I then argued that if we incorporate metalinguistic content into our account, we can handle these cases, implying that we use the word ‘concept’ sometimes to denote something like the meaning of a word in a shared language. The account we have proposed takes phenomenal properties as primitive components of the reference-fixing descriptions for both superficial concepts and essentialist concepts. How deferential or ‘public’ concepts have their content determined has however been left open. Refining an account of this process is what I turn to in the next chapter.
APPENDIX: Phenomenal Concepts and Composition

An argument against the possibility of the kind of phenomenally grounded concepts we have been considering to explain the bulk of our erroneous judgments is that they fail to compose in the way in which concepts should. This is an argument from Fodor that is aimed at ‘recognitional concepts’, of the sort discussed in chapter 2, but that applies just as much to the current proposal.

Providing an explanation for how concepts compose is taken to be a basic desideratum of a theory of concepts. If I possess the concept RED and the concept HOUSE, then I ought to be able to understand what a red house is. Similarly, if I possess the concept BLUE, and the concept CAT, then I ought to be able to understand what a blue cat is. Since one constitutive feature of thought is its productivity – that a finite string of concepts can produce infinitely many claims via functions such as 'and' and 'or' – a basic desideratum of a theory of concepts is to be able to explain this productivity. Since the obvious way to explain productivity is via the compositionality of concepts, a theory of concepts should explain compositionality.

However, so the objection goes, recognitional concepts don't do a good job of explaining compositionality. Take the concepts RED and HAIR, suggests Fodor:

somebody who is able to say whether something is a good instance of HAIR and whether something is a good instance of RED is not necessarily able to recognize a good instances of RED HAIR. Well then, what does “red” contribute to the semantics of “red hair”? Just what you'd suppose: it contributes a reference to the property of redness (as such). It's just that its doing isn't tantamount to, and doesn't entail, its contributing a recognitional capacity for (good instances of) redness. One's recognitional capacity for RED doesn't compose. So one's recognitional capacity for red things is not a satisfier for the concept RED. So not even RED is a recognitional concept. (Fodor, 1998b, 42)

If possessing the concept RED amounted to being able to recognize red things when you see them, and possessing the concept HAIR amounted to being able to recognize hair
when you see it, then by compositionality, possessing a recognitional capacity for each of these ought to provide you with the ability to recognize red hair. But it doesn't, since we may be confused by the claim that red hair is red at all – it’s far from scarlet. So it seems that the concept RED HAIR is not derived from the conjunction of a recognitional capacity for red or a recognitional capacity for hair, and hence that concepts must not be recognitional capacities.

It appears to me that this argument is far too quick. The problem is that words in public languages that have their root in the conjunction of two words do not necessarily derive their meaning from the meaning of their constituent terms as those terms are currently used. Rather, they may derive their meaning from the conjunction of the meanings of their constituent terms at the time the conjunction was coined. How are we to know that the term 'red', to return to the example of 'red hair', did not originally cover a range of colors that it no longer covers, but the expression 'red hair' has retained the meaning of the term 'red' in its older use? This would seem like a very plausible analysis, and the English language is rife with examples of this. Such compounds, which do not straight-forwardly derive their meaning from the meanings of their conjuncts, are known as idioms. Here are some examples, in the first two where the conjuncts have retained their meaning but the meaning of the conjunction has changed, and in the second where the conjunctions have retained their original meaning but the meaning of the conjuncts have changed:
<table>
<thead>
<tr>
<th>Conjunction/Disguised Conjunction</th>
<th>Original Meaning</th>
<th>Current Meaning</th>
<th>Original Meaning of Conjuncts</th>
<th>Current Meaning of Conjuncts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacture</td>
<td>Hand Made</td>
<td>Mass Produced</td>
<td>Hand+Made</td>
<td>Hand+Made</td>
</tr>
<tr>
<td>Awful</td>
<td>Full of awe/ full of wonder</td>
<td>Abhorrent</td>
<td>Awe+Full</td>
<td>Awe+Full</td>
</tr>
<tr>
<td>Blue Heron</td>
<td>Grey Heron</td>
<td>Grey Heron</td>
<td>Grey+Heron</td>
<td>Blue+Heron</td>
</tr>
<tr>
<td>Mincemeat</td>
<td>Minced Food</td>
<td>Minced Food</td>
<td>Minced+Food</td>
<td>Minced+Animal Flesh</td>
</tr>
</tbody>
</table>

**Figure 3: Conjunctions and idioms**

These terms derived their meaning at the time of their original use as a conjunction from the meaning of their conjuncts as those conjuncts were used at the time. A manufactured object is, according to the conjunction of the original meanings – the latin verb *facere* (to perform) and noun *manus* (hand) – one whose construction was performed by hand. At the time the conjunction was coined, anyone who understood *manus* and *facere* would understand what a manufactured object might be. But once the conjunction enters common use, its meaning can depart from the meaning of its conjuncts, given changes in the meaning of either the conjuncts or the conjunction – and the term now means the opposite of hand-made. In the last two cases this is particularly clear. In the last case the meaning of ‘meat’ has changed in the vernacular from meaning ‘food’ to meaning the flesh of an animal, so that the conjunction ‘mincemeat’ has retained the original meaning of ‘meat’, but the use of ‘meat’ has become specialized to pick out animal flesh alone. In the second last case the term ‘blue’ has had an extremely variable history but at times has picked out color shades ranging from green to grey to black. The conjunction that yielded the descriptive name ‘Blue Heron’ retains the original meaning of ‘blue’, while the term ‘blue’ outside of that conjunction does not. Again, anyone who understood the meanings of ‘blue’ and ‘heron’ at the time of the coining of the name would have recognized the bird as falling under the description ‘blue’.
There is every reason to suspect that the term 'red' originally covered a larger range of shades than it is used to cover today, but has ‘specialized’ just as the terms ‘meat’ and ‘blue’ have – we find odd applications of the term ‘red’ in ‘red hair’ (arguably auburn), ‘red brick’ (arguably brown) and ‘red beans’ (arguably purple). The term 'red' used as a predicate now covers one range of shades, but the conjunction 'red hair' has plausibly retained the original broad meaning of the term 'red'. That the pair of words 'red hair' may operate as a conjunction in disguise, so to speak, that is without the words being conjoined on the page, is given further evidence when we consider that 'redhead' is one word. So we can see that the fact that the term in the English language 'redhead' is not necessarily understood by someone who understands the terms 'red' and 'head' does not at all indicate the failure of a recognitional model of concepts.

The point of all of this is that we needn't expect conjunctions or disguised conjunctions in public languages to derive their meanings from the conjunction of the meanings of their conjuncts as those conjuncts are currently used. So evidence to the effect that it is possible to know the meaning of a pair of conjuncts but not know the meaning of the term that derived at some point from their conjunction is not evidence against the compositionality of prototypes or recognizable profiles at all – it might only be evidence that the meaning of conjunctions can depart from the meaning of their conjuncts in a public language, or vice versa.

Fodor's response to claims such as these is that they can't plausibly cover the range of conjunctions that fail to compose if they are conceived on a prototypical model:

Prototype theorists, in their desperation, are sometimes driven to suggest that MALE NURSE, PET FISH, and the like aren't compositional after all, but it's all right that they aren't, since they are idioms. But surely, surely, not. What could be stronger evidence against PET FISH being an idiom or for its being compositional than that it entails PET and FISH and that {PET, FISH} entails it? (Fodor, 1998a, 103)

The weight of evidence seems even-keeled here, however, between supporting the prototype theory and counting against it. The evidence for PET FISH being an idiom, according to the prototype theorist, is simply that good instances of PET FISH don't derive straight-forwardly from good instances of PET and good instances of FISH. The evidence against it being an idiom, according to the atomist, is that it appears to involve the
concepts PET and FISH uncontroversially. But whether it does involve these concepts uncontroversially will, of course, depend on whether it is idiomatic or not. Fodor has more recently recognized that it would in general be extremely difficult to disprove the supposition of idiomaticity in these cases (Fodor and Lepore, 2002, 41). For all we know, it may be the case that public languages are *massively* idiomatic – and given assumptions about recognitional capacities underwriting conceptual capacities, the frequent mismatch between the composition of prototypes and the composition of the meanings of terms in a public language would count as strong evidence for that very conclusion. So we should really be appealing to some sort of independent principle to decide who to believe. I think we have established so far that such an independent principle is available. The appeal to phenomenal properties to determine concept membership explains how we manage to group members of single classes together under one concept, and it also explains why we sometimes erroneously judge of objects that don’t fall under the concept that they do. The rejection of the constitutive link between these observable properties and our conceptual capacities deprives us of these explanations. The weight of argument, therefore, seems to strongly favor the phenomenal properties account.
In the last chapter we identified the role that language can play in thought in underwriting some of the normative constraints that we place on concept application. When we make judgments in terms of the reference of the meanings of terms in a shared language, our judgments about the extension of the terms that play a role in our thoughts can be false. I argued that the best way to think about these extensions is as the contents of what I proposed to call 'public concepts'. What I did not discuss is how these public concepts acquire their contents, which is my aim in this chapter. Obviously, the story we will tell is going to involve language. We think of the extension of a public concept under the sense 'the reference of 't'', where 't' is a term in a language that we share with other speakers. The question we want to ask, then, is what fixes the public meaning of the term, and hence determines the content of the public concept? At this point the current project links up with a strand of thought that was first introduced by Putnam (1975) and Burge (1979) under the phenomenon of what the latter called ‘deference’. These writers proposed that for the meanings of some of the terms in our thoughts and utterances, we ‘defer’ to the knowledge of those more expert than ourselves. The question I want to ask in this chapter is how ‘deference’ works, and therefore exactly what fixes the content of these ‘deferential’ terms.

1. Varieties of Deference

The two standard models of deference in the literature are what I will call expert deference, and typical use deference. I will now examine both, to see if they can give satisfactory answers to our two key questions.
1.1 Expert-Deference

The introduction of the idea of deference to the literature is given in Putnam (1975) and Burge (1979). Putnam realizes that in some cases the superficial concepts that he has of objects in his environment will not be enough to distinguish objects that he can nevertheless distinguish in thought. In most cases, Putnam supposes that the content of his thoughts is determined by facts about objects in his physical environment that are distinguished according to certain stereotypical features. So, his 'water' thoughts are distinct from his 'beer' thoughts because he can superficially distinguish water and beer in terms of certain surface properties. The hidden or essential properties of the classes picked out by their superficial properties then determine the extension of the thoughts that pick out those classes. In some cases, however, Putnam can distinguish classes in thought that he does not have distinct stereotypical ideas for. He can wonder of a tree that he encounters whether it is an Elm or a Beech, without being able to think of any superficial properties of Elms or Beeches that would be sufficient to distinguish the two, or recognize the difference between them. In these cases, Putnam proposed that we have a 'linguistic division of labor', where I defer to the opinion of those more knowledgeable than I for the meaning of my utterances 'Elm' and 'Beech'. The content of my ‘Elm’ thoughts is picked out by whatever the local tree experts tell us Elms are. By thinking of the meaning of ‘Elm’ as ‘whatever the experts around here call ‘Elm’”, I can isolate Elms from Beeches in spite of being unable to do so perceptually.

Burge (1979) further developed Putnam's suggestion, arguing that not just the meaning of my words, but also the content of my thoughts is subject to deference to experts. If I believe I have arthritis in my nose, but my doctor tells me that this is impossible, then I will concede that all along what I was thinking of as ‘arthritis’ was in fact something that could not occur in my nose – thus deferring to the doctor’s views on the contents of my own beliefs and thoughts. It is unclear whether Putnam and Burge think that the opinion of a single expert at some point at the top of a pyramid of more and more knowledgeable individuals determines the meaning of my utterance of expert terms, and the content of my beliefs that involve expert terms. However, this is certainly the idea of ‘deference’ that has made its way into many models, including the model of
concepts discussed in chapter 1. Similar ideas appear in Searle (1983), who discusses how it is possible to 'piggy back' in thought upon the thoughts of others. Searle supposes that in some cases I can identify a particular thinker and 'defer' to his mental states to determine the content of my thoughts about, for example arthritis, by thinking of 'whatever he calls 'arthritis''. Fodor upholds this view, arguing that “If my ELM concept is deferential, that’s because the botanist’s isn’t” (Fodor, 1998a, 154). The assumption in all these accounts is that although deferential thinking certainly takes place, the buck stops somewhere, and with an individual thinker. And this is the model of deference that has made its way into some recent analyses, such as Chalmers’, who adopts Kripke’s example of Leverrier into a case of deference by granting Leverrier’s wife cognitive access to Neptune via the description 'whatever my husband refers to by the term 'Neptune'' (Chalmers, 2002a).

Now expert-deference is in a position to give a compelling answer to the second of our two questions – why do people defer. It is clear that there is an advantage in deferring on the meaning of some term to someone that you have good reason to think is more knowledgeable on the subject than you are. Deferring to your doctor on the meaning of ‘arthritis’, for example, allows you to carry out an informative conversation with your doctor on the question ‘do I have arthritis’. If you have good reason to think your doctor knows more than you do about arthritis, then it is advantageous to believe his answers about whether you have arthritis, and not your own – hence deferring to your doctor on the questions what arthritis actually is. Deferring to your doctor is good for your health.

However, the expert-deference model fails badly when it comes to the second question: how does it work? As we saw above, what all these models have in common is the claim that at some point along a chain of deference, a fact about some individual thinker determines the content of the deferential thinker's thoughts. As we shall now see, we have very good reasons for thinking that such a claim is untenable.

Who are the Experts?
An initial worry for the expert-deference view, is that it seems that it would be necessary for me to know who the experts are in order to defer to their views in the piggy-back form that we might expect our judgment to take. According to Searle and Chalmers, I can defer to the views of another by picking that individual out and thinking ‘whatever he means by ‘t’’. And no doubt I can do this. But this cannot be the only way that I can defer – since the content of my ‘crustacean’ thoughts are deferential, but I have no idea who the crustacean experts are. So I cannot identify the crustacean experts and defer to what they mean by 'crustacean'. This worry can be handled, though, by appealing to chains of expertise: for the term ‘arthritis’, I defer to my neighbour who has been speaking with a doctor, he defers to the doctor, the doctor defers to the Doctor’s Handbook, the author of the handbook defers to the World’s Greatest Arthritis Expert, and so on. In this way, deferring to ‘whatever he means by ‘arthritis’’, where he is my neighbour, may end up being deference to the World’s Greatest Arthritis Expert.

How many Topics have Experts?

A more serious problem for the expert-deference view, however, is that even if we allow chains of deference, given the number of terms on whose meaning we are likely to concede to correction, it does not seem likely that there could be experts for all of these topics. In the case of a term like 'arthritis' or 'crustacean', it seems plausible enough that there are discernible individuals in my community who are more knowledgeable than I on the topic, and to whose judgment I am likely to concede if they were to correct my use of the term, or my beliefs about the extension of the term – my doctor or zookeeper being obvious examples. But concession to correction on our application of terms extends far more widely than to obvious expert terms. I might concede to correction on whether some color counts as 'mauve' – but are there really mauve-experts out there who are deciding what is to count as mauve and what isn't? What about a term like 'shoe'? It is not inconceivable that someone could concede to correction on, say, whether spats count as shoes or not. I might believe they don't, but concede to correction if someone shows me a copy of the Oxford English Dictionary that defines 'shoe' in such a way as to clearly include spats. But of course the authors of the dictionaries are not deciding what counts
as a shoe – they are merely recording standard usage. So who does the Oxford English Dictionary consult – do they call the shoe expert? It seems extremely unlikely that there is one.

*When did the Experts Acquire Their Expertise?*

The biggest difficulty for this view is reminiscent of Quine's (1936) regress charge against conventionalism about linguistic meaning. Even supposing that there are currently experts on various topics, it seems that it cannot be the case that there always have been. The development of a body of expert knowledge on, say, arthritis can only have taken place given a shared understanding of the meaning of the term 'arthritis' that the various experts used to make sure they were all talking about the same thing. Presumably there were correctness conditions for the use of these terms before a body of expert knowledge, therefore, ever arose. The most plausible explanation for how a body of expert knowledge concerning arthritis arose is that a curious individual at some point in medical history decided to investigate the underlying cause of the condition known as 'arthritis', passed on his findings to his students, and bit by bit individuals with greater and greater expertise on arthritis came about. Doctors *can* invent new classificatory terms, but often they don't – rather, they investigate the origins of conditions that are already well-known and whose names can be used correctly or incorrectly. For the reasons we explored in chapter 2, new terms cannot easily be introduced to a language simply on the basis of the intentions of a baptiser, and so it seems far more likely that the experts come *after* the terms, rather than the other way around.

For all of these reasons it seems extremely unlikely that deference to speaker-independent meanings, evidence for which is found in the extremely widespread concessions to correction from other speakers that we are inclined to make, is deference to the views of a single expert or panel of experts. And so the expert-model of deference fails to answer the first question plausibly: how does it work? The story about how expert-deference actually works seems largely fictitious.

**1.2 Typical-Use Deference**
Another approach to the deference puzzle would be to fix the meaning according to the typical use of a term within a linguistic community (Wittgenstein 1953; Sellars, 1969; Lewis, 1969; Loar, 1988; Williamson, 1994). Williamson holds that the meaning of a term is constituted by the statistical average use of the term within a linguistic community. This account is designed to address the problem of vagueness, which I do not address here, but his account could equally well be recruited to address the puzzle of deference. If individuals defer to the meaning of a term as it is used in their community, which meanings they may be ignorant of, then it will be possible to explain how an individual could refer to a color (say 'red'), and nevertheless fail to know what the extension of the term 'red' in fact is (the vagueness problem). This allows us to explain, on Williamson's account, how vague terms acquire their meanings – not from the individuals who use the terms (since those individuals are unable to specify the extension of the terms) but from facts about the use of the terms in the individual's community. This model can also be used to explain the normative constraints we discussed in the last two chapters – I can be trivially wrong about the ‘typical use’ of a term in my community, and so if typical use determines meaning in some important way, I can be wrong about the meaning of a term.

Now the typical-use model answers our first question – how does it work – quite well. At least at a first glance there is nothing mysterious about the proposal that there is a statistical average use for the term ‘red’ for some given community of speakers. To get the statistical average, you simply take the beliefs of all the individual members of that community about the range of shades to which the word ‘red’ applies, and see what range of shades results as the average from among those ranges. Where the typical-use model fares badly, however, is on the question ‘why do we defer’.

**Irrelevance of Typical Use to Conversational Interests**

First, the statistical average use of a term in a community can be seen to have very little relevance to the interests of speakers in a conversation. This can be brought out when we consider the arbitrary nature of the factors that have to be taken into
consideration when computing a statistical average use, a difficulty which has been called the 'arbitrariness' worry (Hawthorne and McGonigal, 2006). Williamson envisages that the exact meaning of a term may be in a constant and chaotic state of flux, according to the changing uses to which a linguistic community will put a term, which can vary “in an unsurveyably chaotic way” (Williamson, 1994, 209). This presents us with some surprising results for the meaning of the terms in a conversation. Suppose I utter the following sentence:

All animals are equal but some animals are more equal than others (Orwell, 1945)

Now suppose that mid-sentence, someone in Australia (who arguably belongs to the same linguistic community to which I belong – of English speakers), uses the term 'animal' such that the statistical average of uses of the term changes from excluding carnivorous plants, to including them. Are we to suppose that the term as I use it at the start of my sentence refers to a set of creatures that does not include carnivorous plants, while the use at the end of the sentence refers to the set that includes these plants? Or what about from the beginning of a conversation about animals, to the end of the same conversation? Or from one conversation before lunch, to another after? The fluctuating statistical average model will seem to allow for such eventualities, which seem extremely implausible. It is not obvious why I would take the beliefs of someone in Australia, whose opinions are for all I know are extremely strange, to be relevant to the meaning of my utterances when I am engaged in conversation with someone in Toronto, to such an extent indeed that his beliefs would affect the meanings of my utterances mid-sentence.

*Arbitrariness of Language Borders*

The arbitrariness worry can also be articulated by considering the lack of clear borders for linguistic communities. Suppose, as is entirely plausible, that some French word is used slightly differently in Alsace-Lorraine to the way it is used in the rest of France. If we fix the borders of the French language west of Alsace-Lorraine, for example, then the use of this word will have one statistically averaged content, but if we fix the border east
of that region it will have another. In most of France, let’s suppose, the word 'fête' is used to mean 'party', but its use excludes funerals. In most of France, let’s say, nobody thinks a funeral is a fête. But now let’s suppose that in Alsace-Lorraine the term is used to refer to both parties and funerals. In that region, the term means 'occasion of honor', and can include birthdays, weddings, funerals, etc. So if the statistical average use of a term in a community determines the term's meaning, then whether the term 'fête' means 'party that includes funerals' or 'party that excludes funerals' will turn on where the national border of France is drawn. That this decision would determine the meaning of the term 'fête' in any given conversation, especially if determined on the basis of the outcome of the Second World War, seems deeply absurd.

*The Problem of Averaging Incommensurables*67

Another objection still is that while the notion of an ‘average use’ seems intelligible for terms with a continuous range of extension, such as color terms68, it’s not clear how it could be calculated for terms whose use extends over a range whose objects are not obviously of the same class. For example, the term ‘meat’, as we explored in the last chapter, was once used to refer to any sort of food, but is now used to refer only to food made of animal flesh. Now suppose in some regions the term is used to refer to any food, and in another it’s used to refer to animal flesh alone. On the typical-use model, we are to find the average use and define that as the deferred-to meaning. But what is the average between, say, meat and all other food? The question seems unintelligible – there is no continuum upon which all food sits. If the sets of objects to which a term is applied are incommensurable – that is, they don’t fall into a continuous set – then there is no way to define what the ‘average food’ in the set might be.

The typical-use theory, therefore, fails on both counts – although more obviously on the question ‘why do I defer’. What do I have to gain from conceding to the statistical average use of a term in a relatively arbitrarily determined linguistic community for the

---

67 I owe this point to Bill Seager.
68 Although there could certainly be color terms with a ‘gappy’ extension – as Horwich proposes to deal (unsuccessfully in my opinion) with the vagueness puzzle (Horwich, 2003); we can stipulate such terms, e.g. ‘scarange’ which we can suppose means scarlet or orange, but none of the shades between these two.
meaning of my utterances mid-sentence? It seems that such concessions would do more to undermine the communicative efficiency of language than to optimize it.

A more convincing theory of meaning-determination, that would avoid the arbitrariness worry, might be a conventionalist account of meaning (Lewis, 1969; Schiffer, 1970), where speakers enter into tacit conventions on the meanings of their utterances. However, as we explored in chapter 3, it is clearly possible to establish linguistic meanings in cases where there are no conventions governing those meanings. To establish the grounds of public meanings, I will turn to what I call a 'participatory deference' account, where individuals simultaneously both defer and contribute to the determination of the meaning of their utterances, in a process of discursive collaboration that is quite close to the conventionalist account, but ultimately does not rely on conventions.

1.3 Participatory Deference

In chapters two and three, we discussed a proposal for what fixes an aspect of meaning that we called the communicative value of a term. So far, that value was modeled on the basis of the beliefs of pairs of speakers engaged in discourse. The crucial variable in this model was not the convention, or private intentions, but speakers shared beliefs. Neither member of a pair of interlocutors is entirely in control of the communicative value of their utterances, and both members need to recognize that establishing a communicative value is a joint venture for them to succeed at all. This has already introduced to our model of language a minimal form of deference. Each speaker is bound to defer to the communicative value of their utterances in the conversation in which they are engaged as at least partly constitutive of the meaning of their utterances in that conversation. Neither speaker can take control of the communicative value of their utterances single-handedly, so they are forced to work with what they’ve got – and concede to the communicative values that it is possible to establish with their current interlocutor. So here we have at least one species of deference.
However, all this does not quite account for the kind of deference to 'public concepts' we are hoping to explain. While the communicative value may be fixed in a collaborative manner, and while neither speaker in the discourse may be in control of the determination of the communicative value of their utterances, it still does not make sense to say that the public meanings to which we defer are determined by the intersection of the beliefs of speakers in conversation. This account would not, after all, account for the fact that we are perfectly happy to say that a given pair of speakers can both be wrong about the meaning of some term they are using. They may be communicating successfully with that term, but it seems that both speakers in the Hunky Dory conversation considered in chapter 3 are wrong about the meaning of ‘Hunky Dory’. So against what standard are we judging the beliefs of these speakers erroneous?

It doesn't seem as though we can revert to the opinions of experts, or a statistical average at this point, for the reasons we have already given. There does seem to be an option, however, of extrapolating out from the model of shared beliefs that we have been discussing so far to communities of speakers. If we take three speakers, for example, the intersection in their sets of beliefs will be distinct from the intersection of the sets of two speakers. With four, we get a different intersection again, and so on. If we proceed in this manner, however, we find ourselves with the same failure to fully account for deference that we have with just two speakers – out of 100 speakers, if the intersection in their sets of beliefs about the meaning of the term 'red' constitutes the 'public meaning' of the term for that group, then nobody can be entirely wrong. But this is not how we analyze utterances for correctness – it is possible to be completely wrong about the meaning of 'arthritis', as someone who believes that a rash is a symptom of arthritis and that it primarily affects bones and not joints seems to be. And so while the intersection of sets of beliefs, established through triangulation, can determine the communicative value of terms for pairs of speakers, it doesn't seem as though it is sufficient to account for our concessions to being completely wrong about the meaning of some terms in our language. So we have made some way in establishing what public concepts might be, but we haven’t quite arrived. To establish the grounds of public concepts proper, I would like to turn now to a recently developed field of inquiry on what has been called ‘public
knowledge’ – and use insights in that field to develop a theory of public concepts, and deference.

2. Public Knowledge

To secure an answer to the question ‘what determines the content of public concepts’, I would like to turn first to the notion of ‘public knowledge’. This relatively recently recognized phenomenon operates in much the same way, I will argue, as public concepts, and recent research on this topic will help us to establish a model for determining the content of the latter.

2.1 Is there public Knowledge?

To begin with, we can fortify our intuitions about whether there is such a thing as public knowledge by looking at some recent studies on intuitions about attributions of knowledge to groups or institutions (Bloom & Veres, 1999; Gilbert, 1992; Huebner in prep.; Kashima et al., 2005; Knobe and Prinz, 2008; Morris et al., 2001; Pettit 2001, 2004; Searle, 1995; Solan, 2005; Tollefsen, 2002; Tuomela, 1995; Velleman, 1997). The kind of attributions I am talking about would include, for example, where we might say that 'NASA knows how to send a rocket to the moon', although we would deny that any individual member of the NASA knows how to send a rocket to the moon. Each individual has their own expertise, and that expertise contributes to the ability of the group, but no individual would single-handedly know how to put all the pieces of the project in place. And in case we might worry that, as an attribution of knowing how, this might not count as an epistemic state, we could rephrase to say that 'NASA knows what is required to send a rocket to the moon', which also seems true, and is clearly an attribution of knowledge-that. Knobe and Prinz (2008), for example, designed a survey that clearly illustrates that people are generally comfortable attributing epistemic states to corporations.
The primary intention of the survey was to see whether the general public have a distinct idea of what phenomenal consciousness is, or whether we ordinarily fail to distinguish between phenomenal states and functional states. With this end, they asked people whether they were comfortable attributing states such as pain, fear, sadness, or joy to groups of agents such as corporations, and then contrasted their replies with replies given to a similar set of questions concerning the attribution of states that are not obviously phenomenal to such groups, such as knowledge, belief, or understanding. The authors found that people are far quicker to attribute states that are not obviously phenomenal to corporations than they are phenomenal states. This is an interesting finding for the question of whether there is a folk psychological notion of phenomenal consciousness and whether it is tied up with notions of physicality, but it also very nicely illustrates the naturalness of attributions of epistemic states to groups:

We began with a study in which subjects were given a list of sentences that ascribed mental states to group agents and then asked whether each of these sentences ‘sounds natural’ or ‘sounds weird.’ Some of the sentences ascribed non-phenomenal states; others ascribed phenomenal states. The two types of sentences were mixed together, and the order of presentation was randomized.

The sentences ascribing non-phenomenal states were:

- Acme Corp. believes that its profit margin will soon increase.
- Acme Corp. intends to release a new product this January.
- Acme Corp. wants to change its corporate image.
- Acme Corp. knows that it can never compete with GenCorp in the pharmaceuticals market.
- Acme Corp. has just decided to adopt a new marketing plan.

The sentences ascribing phenomenal states were:

- Acme Corp. is now experiencing great joy.
- Acme Corp. is getting depressed.
- Acme Corp. is feeling excruciating pain.
- Acme Corp. is experiencing a sudden urge to pursue internet advertising.
- Acme Corp. is now vividly imagining a purple square.

For each sentence, subjects were asked to provide a rating on a scale from 1 (‘sounds weird’) to 7 (‘sounds natural’). The mean ratings were as follows:

*Non-phenomenal states:*
0.6.6: Deciding
0.6.6: Wanting
0.6.3: Intending
0.6.1: Believing
0.5.2: Knowing

Phenomenal states:
0.4.7: Experiencing a sudden urge
0.3.7: Experiencing great joy
0.2.7: Vividly imagining
0.2.5: Getting depressed
0.2.1: Feeling excruciating pain

As the table shows, even the most acceptable phenomenal state was still deemed less acceptable than the least acceptable non-phenomenal state. More generally, there was a statistically significant effect such that people gave lower ratings for the phenomenal states than for the non-phenomenal states (Knobe and Prinz, 2008, 75-76).

These results clearly show that we are at the very least uncomfortable attributing epistemic states to groups or, in this case, corporations: attributions of epistemic states including deciding, wanting, intending, believing, and knowing are generally ranked as highly 'natural' sounding. So the question we are now faced with is how to decide what the groups in fact know.

2.2 What Determines the Contents of Public Knowledge?

The attribution of epistemic states to these groups may, however, be thought to be attributions of epistemic states to individual members of the groups, to whom everyone else in the group defers (Quinton, 1975). This would be in line with the deference to individuals or expert-deference models we considered above. But there are good reasons for thinking, even in tight-knit groups where if there are any experts the members of the group would likely know who they are, that this is not the right analysis. Velleman (1997) argues that while a philosophy department can intend to hire a new professor, no individual member of the department has the authority to hire the professor on their own. Since you can’t intend to do something you can’t do, the intention to hire a new professor cannot be attribute to any one member of the department, but only to the department as a

Since epistemic states such as beliefs are constituted at least in part by concepts, if we are comfortable attributing beliefs to corporations, we ought to safely assume that such attributions tacitly also attribute concept possession to these groups. And so, we have intuitive traction for the idea that there are such things as public concepts, whose content cannot be identified on the basis of facts about individuals considered outside of their relationships to the other members of the groups to which they belong. The next question is how we decide what determines the content of the beliefs of a group, which will lead us to an answer for the question what determines the content of public concepts.

While identifying the belief of a group with the belief of some individual member of the group doesn't seem plausible, neither does identifying that belief as the sum of the beliefs of the members of the group. Summing the beliefs of the members, for example, could result in the attribution of an incoherent set of beliefs: suppose the majority believe that one proposition A is true; and a majority believe that another proposition B is true; but the majority do not believe that both A and B are true. In that case, we will have assigned to the group an incoherent set of beliefs: \{A, B, and ~(A&B)\} (Pettit, 2001). To attribute an epistemic state to an entity, it is surely a precondition that the entity not hold multiple conflicting views simultaneously (at least at a personal level, where those views are explicitly conflicting). And so to explain our practices of attributing epistemic states to groups, it seems that we cannot appeal to an intuition that the epistemic state of the group is the epistemic state of the majority of its members. The members may have conflicting views, which would result in attributing conflicting epistemic states to the group.

Pettit (2001, 2004) argues that best method for determining the contents of group epistemic states is not through a statistical sampling or a majority vote, but through a process of what he calls 'collective reasoning'. He introduces this idea as a way of solving what he calls the 'discursive dilemma' that arises for any group that finds itself having to make a decision as a group. The discursive dilemma is itself a generalization of a more specific problem that arises in jurisprudence, called the doctrinal paradox, a brief look at which will help to clarify the issues at stake.
The doctrinal paradox arose within jurisprudence upon the realization that two equally legal ways for a panel of judges to decide the outcome of some case can produce conflicting verdicts. Pettit gives an example where a panel of three judges has to make several decisions to decide on the outcome of a case. If the majority vote of the judges is used to decide each question separately, and those answers used to decide the outcome of the case, the result will not necessarily be the same as if each judge is allowed to draw their own overall decision on the case, and the majority view of their final conclusions used to decide the case. For example, if a subject is guilty only if he is found to have been both negligent and as having a duty of care toward the plaintiff, then the following matrix represents the conflicting patterns of tallying votes of three judges A, B, and C:

<table>
<thead>
<tr>
<th></th>
<th>Cause of Harm?</th>
<th>Duty of Care?</th>
<th>Guilty?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>B</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>C</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Guilty?</td>
<td>Yes</td>
<td>Yes</td>
<td>↓ No / → Yes</td>
</tr>
</tbody>
</table>

Figure 4: The Discursive Dilemma, Pettit (2004, 168)

If the question of whether negligence (cause of harm) and duty of care apply in the case is decided by taking the majority view on each question, then each will be given the overall answer 'yes', since the majority (2) voted 'yes' on each of the two premises of the question. If those majorities are then tallied to produce a final verdict, the verdict will be 'guilty'. If on the other hand each judge draws a conclusion based on his independent assessment of the negligence and duty of care questions, two judges will conclude that the defendant is not guilty, and so the final verdict will be 'not guilty'.

The judges in the case above have two options. They can adopt what Pettit calls the 'premise based' strategy for determining the outcome of their deliberation, or they can adopt the 'conclusion based' strategy. The premise based strategy is to calculate the conclusion on the basis of their weighted consensus on the premises. The conclusion
based strategy is, contrarily, to calculate the conclusion on the basis of their consensus on the conclusion irrespective of their views about the premises. The former strategy will yield a potentially inconsistent set of commitments on the part of the group: the weight of votes on whether the premises hold says they do, but the conclusion of 'not guilty' denies that they do. The latter strategy, on the other hand, runs the risk of collectively endorsing a conclusion that a majority of the participants as individuals may reject. Pettit holds that groups that adopt the first strategy deserve to be properly thought of as ontological subjects or persons. I will first discuss Pettit's reasoning, and then argue that his model of collective reason provides us with the grounds for a notion of deference to public meanings that is more plausible than the experts or typicality models.

As a criterion for allowing a collective to be attributed intentional states, Pettit argues, the collective needs to be such as to be assessable for rational consistency, and display the sort of 'rational unity required of an intentional subject' (Pettit, 2004, 181). As an upshot of this criterion on being taken as an intentional agent, the collective, according to Pettit, is going to have to prefer the premise-based process of deliberation over the conclusion-based process. By preferring the premise based process, the collective risks committing itself as a group to conclusions that possibly a majority of its members disagree with, but by preferring the conclusion based process, the collective presents itself as endorsing conclusions that follow from premises that it has, as a collective, rejected. The latter procedure would involve presenting the group as being committed to an invalid argument. In order to pursue reasoning at the level of the collective, then, the collective will have to endorse the premise-based approach, since it is only on the premise-based approach that the group can be considered rational as a group.

A Personal Incentive to Follow Premise Based Approach

Now Pettit's argument is simply that if a group adopts a process of decision making that conforms to what we can regard as a collective rationality at the expense of the beliefs of its individuals, then it can be properly regarded as a collectively rational entity, or even an intentional agent. Pettit argues specifically that if a collective wants to be taken seriously as rational by the public (say, a political party), then it ought to prefer the
premise-based process. But we can see that the process identified by Pettit has more practical advantages than this still. For every member of the group, the process identified by Pettit is more amenable as a means of arriving at a decision representative of his views, than the conclusion based process. The reason is that because this process is more closely analogous to the process of reasoning of an individual agent, it can be more readily accepted as a process that an individual would accept as rational. If as a member of a group, I am to be bound by the conclusion of that group, then I want the conclusion that the group arrives at to be established in a way that I would consider to be ideally rational. So I ought to prefer the premise-based process simply on the basis of desiring the group to which I belong to act rationally. This process more closely matches my interests as a rational agent, and so will be a process whose outcome, which may involve my having to concede my own view to views that others prefer, will nevertheless be in accordance with my own rational standards.

The premise-based process of collective reasoning thus has two advantages for a group: the first is that it mirrors the process of reasoning an individual would undertake, and hence will be preferred by the individual members of the group according to their own standards of rationality; second, it does not result in the production of inconsistent or contradictory outcomes, unlike the majority-vote model. What I will propose next is that something similar to this process underlies our concession to the views of groups about the meanings of linguistic terms.

3. Public Concepts

So far we have established a plausible means for a group to arrive at a conclusion. It involves, particularly, that some members of the group concede to the majority view on their beliefs about certain background beliefs – the ‘premises’ of the argument considered above. This allows that the group arrive at a unified conclusion, in an ideally rational way. And so the notion of public knowledge is accompanied by a plausible account of
deference: within a group, the members defer to the conclusion that their group would
arrive at given an ideally rational process of group deliberation. Much of the time, we can
suppose, such ideal processes will not actually be carried out. NASA might never sit
down and figure out, as a group, exactly what the rules are for sending a rocket to the
moon (this seems unlikely in the case of that particular project, but clearly it is not
impossible). But the members of NASA can recognize that there is a fact of the matter
about what that knowledge is, even before it is ever made available for scrutiny through a
process of collective deliberation. So the group can be thought of as being deferential to
that collective knowledge, even though no individual knows what the knowledge is. Their
derference is to what the group would agree upon if they in fact sat down and deliberated
upon the question.

So much for deference to public knowledge. We can now consider how this
model of deference might provide us with an explanation for the determination of the
public meanings of terms in our linguistic community. We can note that at the very least
we already have a good reason for expecting that such processes may also determine
contents for public concepts – if we can attribute beliefs to groups, then since concepts
are the constituents of beliefs, we must be able to attribute concepts to groups too. My
proposal for what constitutes the content of these concepts is a combination of the notion
of triangulation discussed in section 2 and the notion of group knowledge discussed so far
in this section. Essentially, my proposal is that the contents of public concepts are
determined as the contents that the members of a community would agree on assigning to
the concept-terms in their shared language under a process of collective deliberation. To
build up to this model, I will revisit the notion of triangulation, and marry it to a model of
group deliberation to arrive at the content of a public concept for the concept MAUVE.

Let us suppose we begin with a pair of speakers, coordinating around a series of
color patches to arrive at a communicative value for the term 'mauve', as discussed
already in chapter 3. One speaker utters the term 'mauve' every time a shade that falls
under her MAUVE concept appears. The other speaker attempts to recognize what each of
the color patches that stimulate the other speaker's utterance 'mauve' have in common,
and acquires a belief about the range of shades to which 'red' applies. Now suppose the
second speaker is naturally inclined to group the range of shades that she has seen the
first speaker call 'mauve' under a set of shades that is slightly different from the set that the first speaker is inclined to group them under. In that case, each speaker will have a distinct MAUVE concept, and the intersection of the sets constitutes the communicative value for the term ‘mauve’ in discourse between these two speakers (fig. 5).

Figure 5: Case 1 - Communicative Value of 'Mauve' for speakers A and B

To the pair, we now add a third speaker. Very likely, the third speaker's dispositions to group the shades will more closely match one than the other of the first two speaker's dispositions (fig. 6). In this case, the intersection of the three sets – what all three agree on as falling under the extension of ‘mauve’ will be drastically diminished by the addition of a third speaker.

Figure 6: Case 2 - Weaker Communicative Value of 'Mauve' for speakers A, B and C
And, as more speakers are added to the group, the intersection of the sets will decrease further, to the point where a single intersection may disappear, and the term 'mauve' will cease to have a single communicative value for the entire group (fig. 7).

**Figure 7:** Case 3: Given sufficient members of the group, without concessions between speakers a single communicative value quickly disappears. Here, we have a single value for the group \{A, B, C\}, and another for the group \{B, C, D\} – but none for \{A, B, C, D\}.

We now have various puzzles that the group is faced with, mirroring the difficulties faced by the groups discussed by Pettit. Just like the panel of judges, it is in the interest of the group that they arrive at a single meaning for the term ‘mauve’, in order that the term have the most effective communicative value possible. Certainly, no member of the group in figure 6 is going to be able to use the term ‘mauve’ to communicate about any shade simultaneously with all other members of the group, since there will always be at least one member, for any shade, who does not share a belief with the speaker that the term ‘mauve’ picks out that shade.

However, there are various strategies for securing a possible meaning that we can see already will be problematic. It will not work to identify the meaning of the term for the group as the meaning agreed on by the majority (in cases 2 and 3 at least), since here we will get the same results that Pettit identified for majority polls to determine the
beliefs of groups. In case 3, for example, three out of four believe 1) $A \cap B \cap C$ is mauve, and three out of four think that 2) $B \cap C \cap D$ is mauve; but there is no majority that believes that $(A \cap B \cap C & B \cap C \cap D)$ is mauve, and so the majority strategy would attribute to the group inconsistent beliefs about the meaning of the term.

What we might explore as an alternative, then, is whether a process of collective reasoning will secure a meaning for the term in a more agreeable manner. Certainly, the members of the group in case 3 are faced with a similar difficulty as the panel of judges considered above. To make the analogy clearer, take the speakers A, B and C in case 2 considered above, and suppose the question they have to answer is whether the term ‘mauve’ applies to the all the shades picked out by B, including $A \cap B$ and $B \cap C$. The following pattern of questions mirrors the questions deliberated on by the panel of judges:

<table>
<thead>
<tr>
<th></th>
<th>$A \cap B$?</th>
<th>$B \cap C$?</th>
<th>$A \cap B &amp; B \cap C$?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>B</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>C</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>$A \cap B &amp; B \cap C$?</td>
<td>Yes</td>
<td>Yes</td>
<td>↓ No / → Yes</td>
</tr>
</tbody>
</table>

*Figure 8: Discursive Dilemma for the Meaning of ‘Mauve’*

If the conclusion is derived from the individual speaker’s conclusions on the basis of their own premises, then the conclusion will be that the term does not apply to the full range of shades. If the conclusion is derived from the majority vote on each of the premises, on the other hand, the conclusion will be that the term does apply to the full range of shades picked out by B, and A and C will have to alter their beliefs about the meaning of the term accordingly, to approach a match with the conclusion arrived at (figure 9).
As Pettit has argued, the premise-based strategy will most closely match the process of deliberation that an individual intentional agent would undertake, and so if the group wants to be considered as an intentional agent, this is the process it should adopt. However, as I pointed out above, this also has the advantage that each individual member will be more amenable to this process of reasoning – because it matches the process they would undertake themselves (i.e. settling on the premises of their argument before reaching a conclusion). This will be of significance when we return to our two key questions that any model of deference must be able to answer.

What this provides us with is a model for how groups might make decisions on the meaning of linguistic terms in the absence of salient experts, and in such a way that would allow that the opinions of all members of the group be taken into consideration. It certainly will result in some speakers having to make greater concessions than others, but if all speakers accept that the process of deliberation most closely matches their own processes of ideally rational deliberation, then all speakers will have an interest in this procedure being carried out.

We might now expect that the results that such a procedure would yield, carried out for all of the members of a community, is what individual speakers within the community defer to. The results are not necessarily ever computed – no individual in a
community may ever discover what such a procedure of collective reasoning would yield as the maximally efficient intension for any given term, just as NASA might never sit down and draw up a manual to determine what their group would decide on as the best way to get a rocket to the moon. But pressure from within the community to maximize the efficiency of the terms in their shared language ought to be enough to get the individual members in that community to concede to whatever is the maximum consensus among their peers, just as pressure from within NASA to actually get a rocket to the moon will force individuals within the company to adopt decisions that ultimately produce an effective procedure for getting a rocket to the moon, and which may involve individual members of the company making many concessions to the greater agreement between other members of the group along with way.

Participatory Deference and the Desiderata of a Theory of Deference

What this gives us is a picture of deference that meets both of our basic criteria without any mysterious postulates: an explanation for how the model works, and an explanation for why individuals would defer under the model. It answers the first question by appealing to a process that we can already know takes place among panels of judges faced with arriving at a consensus on some question, and one that we thereby know is possible. And it answers the second question by identifying an advantage that such deference holds for a community – that of maximizing the communicative value of the term for the community, in a way that respects the rational interests of every participant, by most closely mirroring their own rational processes.

What I would like to suggest is that it is such a notion that underlies the idea of a public concept – of ‘the’ concept MAUVE, as opposed to any individual’s understanding of what counts as mauve. The public concept MAUVE is, therefore, the range of shades that a community would agree on, through the process of concessions outlined above, as the reference of the term ‘mauve’.

The experts model fared well on the question of motive, but failed to answer the first question, given the notable absence of experts for many deferential terms; and the statistical average model, although it gave a clear account of what we are deferring to,
failed to answer the question of motive, given a lack of any apparent advantage in deferring to a wildly fluctuating statistical average mid-conversation. The participatory deference account, on the other hand, provides answers to both of these questions.

**Conclusion**

In this chapter we proposed a model for how the communicative value of terms is established, and by extension how the public meanings of terms, or ‘public concepts’, are established. This provided us with a model of deference that relies neither on the implausibility of appeals to either experts or to statistical averages as the determiners of the content of ‘public meanings’. Rather, on the current model the individual defers to whatever meaning for a term would be settled on if a process of ideal collective reasoning were to be carried out by the community. Hints at the direction that meaning would take are available to individual members of the community by observing the linguistic use of the members of the community they are immediately surrounded by, and identifying the greatest consensus in use among that group. These practices produce meanings for terms that are independent of the beliefs of any member of the group, but that are nevertheless determinately fixed by facts about the group taken as a whole. In the next chapter, I will take an overview of the semantic model that has emerged at this point, and deal with further objections that may arise for that model.
APPENDIX: DEFINING LINGUISTIC COMMUNITIES

Before moving on, however, we should address the arbitrariness worry raised above against the ‘typical use’ model. Since one version of the arbitrariness worry was that it seems to demand that we draw boundaries around linguistic communities in a manner that did not have any obvious connection to the interests of the speakers in any conversation, this may appear to arise for the current model also. Answering the question ‘what would my linguistic community agree on if they could engage in an ideally rational process of deliberation’ demands that we draw a line around the linguistic community appealed to. Unlike the typical use model, however, the current model gives us some reason for drawing a line around communities at particular points. On the current model, the goal is to maximize communicative value for a term in a linguistic community that extends from the speaker. This goal points toward a demarcation line for distinguishing communities: linguistic communities will be distinguished by balancing out the amount of concessions it would take from individual speakers to reach a point of convergence with a group on the meaning of some terms, against the amount of work it would take for them to reach the point of convergence in another linguistic community.

So, for example, if speakers in the Alsace-Lorraine region have more linguistic beliefs in common with French speakers than German speakers, then a grouping of the linguistic community with the intention of maximizing communicative value for their linguistic terms would place those speakers in the French community of speakers. If on the other hand the speakers in that region use certain terms, as a group, with a fluency that would be undermined were the opinions of all the other speakers in France taken into account, then these particular terms should be grouped under a sub-grouping for speakers in that region alone. In other words, if it would take less work (in the form of concessions from individual speakers) for the group of speakers in the Alsace-Lorraine region to converge on a set of beliefs on the meaning of any term than it would for those speakers to concede to the beliefs of a greater number of speakers in France, then the Alsace-Lorraine community will form a distinct linguistic entity. What we have here is a balancing out of the interest of optimizing communicative value against the amount of work it will take to reach that goal. It will take less work to accept that there are two
separate meanings of ‘fête’ between the speakers in most of France, and those in Alsace-Lorraine, than it would for the entire community in Alsace-Lorraine to relearn the term according to the more widespread use. Similarly, it will take less work to accept that French and German are two distinct languages than it would to attempt to get all speakers in both countries to conform to an amalgam of both languages.
In this last chapter I want to discuss the overall semantic framework that has emerged at this point. In section 1 I discuss the model of conceptual content proposed here, and the current model's treatment of the 'qua problem' and the problem of error for conceptual content. In section 2 I turn to the current model’s treatment of linguistic meaning. I discuss again how the qua problem for linguistic meaning is distinct from the qua problem for conceptual content, and how the puzzle of conceptual error is also ambiguous between a puzzle about how individuals’ concepts are applied erroneously and how individuals apply and have thoughts that involve what I called 'public concepts'. Finally, I outline the overall structure of thoughts taking into account all three dimensions now articulated – internal, external, and social.

1. Referential Terms and Conceptual Content

In the first chapter I explored the problem of squaring the content of our ‘epistemic thoughts’ about the objects in our environment with the apparently conflicting content of ‘modal thoughts’ about those same objects. I argued in favor of a two-dimensional descriptivist account of the content of these terms, whose reference is grounded in reference-fixing descriptions. These reference fixing descriptions determine two things: the a priori knowledge we have of the actual reference of the term – for example that it is watery-looking; and the reference of the term – what the term in fact picks out.

This establishes two factors that contribute to meaning – the mind-dependent content of the reference-fixing descriptions, which can pick out different objects in
different possible worlds – and the mind-independent objects that determine what reference the reference-fixing description picks out when it is used to ground a term in the actual world. I will now call these ‘I-content’ and ‘E-content’, for internal and external respectively. In the second chapter I explored a problem facing this account – that the phenomenal ‘snapshots’ that descriptivism has to be restricted to in order to avoid the epistemic problem are not sufficient to distinguish between the various possible objects of reference that satisfy these descriptions. To do this work some authors have appealed to a recognitional capacity on the part of the thinker for the object in question. However, appealing to such a capacity violates the Generality Constraint that I argued is a precondition on attributing a thought about a particular object to a thinker, and anyhow potentially succumbs to a similar puzzle of underdetermination given the multiple recognitional capacities that can be triggered in an encounter with any object. As an alternative, I proposed that we write ‘governing laws’ into the reference-fixing descriptions, that will do the same work as a recognitional capacity but that operate at a personal level and hence satisfy the Generality Constraint.

In chapter 3, I explored a further problem still for the account of reference-fixing proposed. The problem is that we seem to be able to apply referring terms erroneously. Like the qua problem considered above, this problem has been typically considered as a problem for causal theories of reference, but I argued that it is just as much a problem for a descriptive-phenomenal account, since the descriptive-phenomenal account relies on our having a disposition to see certain stimuli as phenomenally similar, or having a disposition to apply the same phenomenal terms to the same causal stimuli. A problem for a dispositional account of content therefore becomes a problem for the phenomenal-descriptive account. I argued that standard explanations of the problem of error are not convincing. As an alternative way of looking at the problem, I argued that once we realize that the judgments about which we can be wrong are invariably made on the basis of the perception of certain appearances to hypotheses about the object underlying those appearances, we can see that in fact we can preserve the ‘dispositional’ link between the reference of our concepts and the tokening of our concepts. Most judgments are not ‘categorical’ or content-determining of our concepts, but merely hypothetical. This leaves us with a picture of the content of the referring terms in our thoughts that fixes reference
via phenomenal descriptions plus ‘governing laws’ that determine which object those phenomenal descriptions pick out. Those reference-fixing acts determine the content of the concept, in judgments that I called ‘categorical’. Subsequent judgments often involve a tokening of the concept on occasions where the reference of the concept is not present; however, as with most judgments subsequent to a content-determining original judgment, these are not relevant to the content of the concept since they are not tokened directly by an external stimulus, but only indirectly via the tokening of another concept.

Both the discussion of the ‘qua problem’ and the discussion of conceptual error, however, left notable phenomena unaccounted for. In the case of the qua problem, the standard analyses, and my proposed alternative, touched only on the question of the determination of the reference of the baptizer’s thoughts, but offered no explanation for how the speaker’s thoughts confer his utterance with a meaning that is taken up by his community. And in the case of conceptual error, I argued that many cases cannot be analyzed in terms of a thinker’s erroneous application of a concept he has an independent grasp of, but must appeal to facts about the thinker’s linguistic peers to be explained. This takes us to the question of linguistic meaning.

2. Shared Beliefs and Linguistic Content

I raised the possibility that the shared beliefs of communities or groups of speakers might determine an aspect of linguistic meaning in chapter 2, where I discussed the qua problem. Since the determination of a ‘baptizer’’s thoughts was not enough to guarantee the conferring of a linguistic meaning on his utterances, I proposed that the constraint that determined what meaning a new term introduced to a language would take on is that of the shared beliefs of the baptizer and his audience. Indeed, if the baptizer attempted to introduce a new term to an audience who acquired shared beliefs that contradicted the intensions or beliefs of the baptizer, it appeared that it was these shared beliefs that would in fact fix the meaning of the term. I argued that what was being set in place by these
shared beliefs was the communicative value of the newly introduced term, and that only given shared beliefs would any communicative value be conferred upon a term.

In chapter 3 I developed this idea of shared beliefs as the ground of linguistic meaning further by exploring the debate between those who hold that speaker intentions act as the ground of linguistic meaning, and those who hold that public conventions act as that ground. I argued that the intentionalist model is unconvincing because it gives us no plausible account of the phenomenon of making linguistic mistakes, and in fact ultimately reduces an individual’s linguistic meanings simply to their thoughts, and hence reduces language to thinking. I then argued that the conventionalist model is also implausible, the most significant reason for which being that linguistic meaning can come about accidentally, with no conventions back it up. I then appealed once again to the ‘shared beliefs’ proposal, and argued that what is really grounding linguistic meaning is the extent to which speakers share beliefs about the meaning of certain terms, or more precisely by the intersection of speakers’ beliefs about the meaning of linguistic types.

In chapter 4, I argued that various cases often taken to be key examples of conceptual error trade on an ambiguity in the notion of ‘concept’, and can only be understood as cases of deference to the meanings of terms in a public language. I proposed that we think of these meanings as what we might call ‘public concepts’, since they are often spoken of as concepts, as in when we say that an individual fails to grasp a concept just when they fail to understand the meaning of a term in a shared language. In chapter 5 I looked at the question of the determination of the content of public concepts more closely. I argued that the standard models of deference in the literature are unsound, given that neither can give convincing answers to both of two simple questions about the models: how they work, and why people would defer to the meanings these models predict. I then proposed that a more promising model lay in developing the notion of communicative value that I had explored in chapters 2 and 3. In order to extend this notion to groups of more than two, and to decide on what the content of a public concept was when plausibly no two people’s beliefs about the concept would overlap precisely, I proposed to draw on the notion of public knowledge and the mechanisms that allow groups to reach decisions. I concluded that the ‘public meaning’ of a term, or the public concept the term expressed, would be that on which the community would agree having
undertaken an ideally rational process of deliberation on the question. This presented the notion of linguistic meaning, or as I have proposed to call it ‘public content’ as an emergent phenomenon constituted by the interaction of groups of people. Since on the model presented, plausibly no single thinker would possess the precise belief about the meaning of the term under consideration, the model if successful would show that linguistic meanings are irreducible to the beliefs of any speakers taken in isolation.

3. Three Dimensional Semantics

The framework that I have built up to this point suggests that meaning has three major components. The first component is the Fregean/internalist component. This component predicts the psychological results of, for example, the discovery of coextensive identities such as ‘Jekyll is Hyde!’ As with most two-dimensional theories, this component is assumed to be determined by facts internal to individual thinkers/speakers. On the current account I have assumed such components to be individuated as rigidified descriptions, which determine what I will call ‘red’ or ‘the morning star’ depending on what world I’m in. This aspect of referring terms also gives us a certain amount of information about what those terms would refer to in different possible worlds – in any possible world, those terms would refer to things with the phenomenal properties picked out by the epistemic profile. And finally, this reference-fixing description determines the a priori knowledge I have of the reference, explaining the role of the term in a priori reasoning and satisfying Evans’ Generality Constraint.

The second component is the Millian referentialist component – the component whose content is determined by the context or environment of the thinker. In a world where the chemical structure of water is XYZ, my ‘water’ utterances and thoughts have the content XYZ. My uses of the term ‘I’ refer to me, but your uses of the term ‘I’ refer to you. The descriptive content of these terms – what we can know about their referents in the world in which we are using them a priori - is plausibly identical, but differences in ‘Millian’ content arise as a result of differences in the context reference-fixing. So the extension is different between the utterances/thoughts, while the Fregean component or
the intension remains the same. The Millian component is eligible for status as a second component of content because it can change while the Fregean component remains the same. So two thinkers in different contexts can have distinct Millian components to their thoughts, with identical Fregean components.

I would now propose to add to this a third component which is the social, or what we might call ‘Burgean’ component. This is not reducible to facts about individual speakers, but as we have explored in the last four chapters, can plausibly only be identified by paying attention to groups of speakers. I have argued that this content can best be understood as what speakers either do agree on, or would agree on through a process of collective reasoning. Typically this component is thought to subsumed under the ‘Millian’ component, and so not to warrant classification as a third ‘dimension’ or component of content. For example, the Jackson-Chalmers framework takes deferential content to be an aspect of my environment, and so to belong to the ‘second dimension’ of meaning.

However I think this ‘social meaning’ warrants consideration as a distinct component for the following reason. The first dimension of content is intensional and internal, as we saw. The second dimension, on the other hand, is neither intensional nor internal. The Millian component of the meaning of ‘water’ in an XYZ world is XYZ itself – that is, a substance and not an intensional function that picks out a substance. In a world where cars and trucks are robots in disguise, the Millian component of ‘transportation vehicle’ is robots. The Millian component is the robots themselves – not an intensional function that picks out robots.

The socially determined component of content is like the Millian component insofar as it is not reducible to facts about individual speakers. It is possible that you cannot find out what the meaning of ‘shoe’ is in English by taking any one speaker and looking inside her brain. However, the social component is also like the Fregean component insofar as it is intensional. The social component of my ‘shoe’ thoughts is a function determined by what myself and my interlocutor agree the word ‘shoe’ picks out in a dialogue, and determined by what my larger community would agree the word picks out if they could sit down and talk about it. So the social component of my ‘shoe’ thoughts is not a set of objects – it is not extensional, but rather intensional. The social
component therefore sits comfortably neither on the Fregean side of the components of content, nor on the Millian side. In the figure 1, I line up the Fregean, Millian, and Burgean content comparatively, calling them I-content (for intensional-internal), E-content (for extensional-external), and S-content (for shared or social, which is intensional but external) (fig 1.).

<table>
<thead>
<tr>
<th>Component</th>
<th>Intensional/Extensional</th>
<th>Internal/External</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Component</td>
<td>Intensional</td>
<td>Internal</td>
</tr>
<tr>
<td>E Component</td>
<td>Extensional</td>
<td>External</td>
</tr>
<tr>
<td>S Component</td>
<td>Intensional</td>
<td>External</td>
</tr>
</tbody>
</table>

Figure 10: Three Components of Content

Another way to think of the difference is as follows. Millian or extensional content is the kind of thing we have thoughts about - not the means by which we go about having those thoughts. Fregean content on the other hand, at least in first order thoughts, is the component of content by means of which our thoughts pick out their referents. Burgean content is like Fregean content in this respect – it is a means by which our thoughts pick out the referents that they do. When I am having ‘shoe’ thoughts, I am thinking about the class of things that the word ‘shoe’ picks out as it is used by the English speaking community – which is to say the class that the English speaking community would agree on if they could sit down and talk about it. However, the social component of my thought is not the class ‘shoe’ – rather, it is the function that picks out that class. My ‘shoe’ thoughts pick out the class of shoes by means of the linguistic term ‘shoe’ that is playing a role in my thought. Now it’s not quite the same as Fregean content insofar as the Burgean content involves a double-quotation in the metasemantic description, while the Fregean content only involves a single quotation. That is, the metasemantic description for a non-deferential term involves only a single quotation, of the form ‘the reference of ‘t’’, where ‘t’ is a referential term epistemically individuated. The metasemantic description for a deferential term, on the other hand, involves a double quotation, in the
form ‘the reference of ‘the reference of ‘t”’. The embedded referential term picks out the intensional function that is determined by my community’s beliefs about the meaning of certain shared linguistic terms. Nevertheless, it is still an intensional function that is the object of my thought, rather than an extension. And so, as natural classes, it seems we have three distinct components to thought.

Not all of our thoughts will have three components. Some thoughts will have only two components – Fregean and Millian. These are, of course, nonlinguistic thoughts. Thoughts with the general form ‘the reference of ‘t”, where ‘t’ does not pick out the reference of a linguistic term, will have a Fregean and Millian component. For example ‘the reference of ‘pain”, where the term ‘pain’ picks out a set of phenomenally distinguishable experiences, has only a Fregean and Millian component\(^69\). But many of our thoughts will have three components. The Fregean component of a linguistically mediated thought about, for example, shoes, will have the form ‘the reference of ‘the reference of ‘shoes”’; the Burgean component will be the function determined by the description the members of my community would agree is appropriate to the term ‘shoes’ through an ideal process of collective reasoning; and the Millian component will be the objects that fall into that class themselves: shoes.

The final position holds, then, that the content of concept terms will have the general form:

\[ \text{The } \phi \text{ that obeys conditional governing laws } L \]

While the communicative value of some linguistic term ‘t’ for speaker S and audience H will then be fixed as \( P \) iff:

\[ S \text{ believes } (\text{Ref } t = P) \text{ and } H \text{ believes } (S \text{ believes } (\text{Ref } t = P)) \]

And the public content of a term ‘t’ will be fixed as \( P \) for a community C under an ideal process of collective reasoning \( R \) iff:

\(^69\) Other thoughts may only have a Fregean component, and no Millian component – such as thoughts about non-existent or impossible entities. This raises the question of fictional names and empty names, which I don’t have space to explore.
C believes \( r (Ref \ 't' = P) \)
Conclusion

The upshot of the considerations raised in this dissertation is that much of the content of our thoughts is not determined privately, or even by the relations between individual thinkers and the noncognitive facts about their environments, but rather can only be understood as the product of, as we might say, relations between minds. The content of thoughts that feature terms that are determined in this way can be accessed by individuals through discourse with their peers on what they agree on, or even what they would agree on given enough time to discuss the matter. Shared languages mediate such discursions, and the terms in these languages are used to keep track of the resulting ‘public content’.

Over the course of the dissertation I considered four areas in which current models of content fall short of adequately accounting for the determination of content. These were the introduction of new terms to a language, the determination of linguistic meaning in the mature stages of a linguistic tradition, the determination of conceptual content given the phenomenon of conceptual error, and the explanation of deferential content. In each of these areas an appeal to various aspects of what we have called public content bolsters our understanding of how these phenomena arise.

Allowing that the intensional content of some of our thoughts is governed by facts about pairs or groups of speakers makes sense out of, rather than making mysterious, the apparent productivity of conceptual analysis (Cf. Moore’s ‘paradox of analysis’, 1953). It explains how, even if the content of our own thoughts is transparently accessible to us given a certain amount of a priori reflection, there seem to be cases of conceptual analysis that are apparently unending explorations. We are still disputing the nature of justice and the good after thousands of years. How can this be? If many of our concepts are not in fact private, but public, then this open-endedness makes sense. Given that many concepts we employ in thought may be functions determined by facts about groups of speakers, and that these groups are constantly changing, the project of conceptual analysis is likely to be as open-ended as any other political project.

The picture of conceptual content that we are left with, as a phenomenon that is in some crucial senses independent of facts about any speakers taken in isolation, takes pairs
or groups of speakers as the basic units in the determination of the content of much of our thinking. The public or social aspect of meaning has played a role of growing importance in thinking about these topics for some time now, particularly since the work of such writers as Wittgenstein, Burge and Davidson. Hopefully the current project lends some further definition to the form that this aspect of our conceptual lives may in fact take.
Bibliography


Bach, K. (2002). ‘Giorgione was so-Called Because of His Name.’ Philosophical Perspectives 16 (s16): 73-103.


Dretske, F. (1994). ‘If you can’t make one, you don’t know how it works.’ Midwest Studies in Philosophy. XIX: 468-482.


Dupré, J. (1981). ‘Natural Kinds and Biological Taxa.’ The Philosophical Review 90


Tollefsen, D. (2002). ‘Organizations as True Believers.’ *Journal of Social Philosophy* 33:


