Knowledge and Content:
A Theory of Interpretation

Alexander Iain Siantonas

The University of Leeds

School of Philosophy, Religion, and History of Science

January 2023

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

I would like above all to thank Robbie Williams, my primary supervisor, who has supported this project so well, from my first enquiries to him through to the day of submission. Thanks also to John Divers, who was my initial secondary supervisor, as well as Andy Peet, who took to the role superbly. Thanks to everyone at the department, PGRs and staff, who helped me along the way, and thanks especially to my proof-reader and unofficial sponsor, Moira Donald (PhD, University of Leeds).

Abstract

This thesis proposes a new theory of content: optimizing dispositions to know. According to this theory, the correct interpretation of an agent is that on which they are best disposed to know. It is a development of the Interpretationist tradition, surveyed in Chapter 1, and especially of the recent work of Robert Williams, discussed in Chapter 2, and Timothy Williamson, discussed in Chapter 3. Chapter 4 explains the theory and argues that it combines the strengths and mitigates the weaknesses found in the ideas of Williams and Williamson. Chapter 5 explains how it delivers plausible verdicts across a range of edge cases, including BIV cases (5.A), Swampman cases (5.B), and a new 'Vatbrain' case combining features from both of these more traditional examples (5.C). Finally, Chapter 6 argues that focusing only on belief, and not action, leaves the theory at no disadvantage against rationality-maximizing views, either in general (6.A) or with respect to moral terms and concepts (6.B).

Contents

Introduction

The goal of this thesis is to develop a theory of content. That is, it tries to give an account of what words and thoughts mean. When I, for instance, say the word ‘table’, I mean *tables*, the large furniture items composed of a board suspended on legs. Those tables are the content of the word I say. I could also think of tables, in which case I will employ a concept whose content is also tables. My task is to explain what makes it the case that the word and the concept have the specific content that they do: why they mean, or refer to, *tables*. This question is supposed to be illustrative: the overarching question is why words and concepts *in general* have the content that they do. This is not a question about etymology, but about meta-semantics. Compare asking what makes it the case that the United Kingdom is a state. This question is not answered by a historical story, in which characters called Alfred and James and Anne feature prominently. This question is answered by a story about political philosophy, in which we specify the characteristics by which something counts as a state, and how the United Kingdom exhibits those characteristics. I am asking what are the characteristics by which a word or concept, or a whole system of words or concepts, counts as possessing certain content.

I approach this question from within the tradition of *Interpretationism*. I will shortly explain Interpretationism in more detail as part of this introduction, but at a first approximation, the Intepretationist approaches meaning by looking at the agent who speaks or thinks and asking what interpretation of that agent - what assignment of content to their words or concepts - makes most sense. This is often described as applying a *Principle of Charity* to an agent: the more sense our interpretation makes of them, the more charitable we are being to them. I start by reviewing the history of such Principles of Charity, drawing the lesson that interpretation should be epistemic: that is, we make most sense of an agent, when we interpret them as *believing well*, which is not necessarily the same thing as *believing truly*.

Accordingly, I examine two more recent epistemic Principles of Charity. Robert Williams holds that we should interpret agents as maximally rational,[1] while Timothy Wiliamson holds that we should interpret them as maximally knowledgeable.[2] While Williams tries to remain mostly neutral about what the best epistemology is, and so what maximum rationality amounts to in detail, Williamson's theory assumes the knowledge first epistemology he champions. On this view, knowledge is the fundamental concern of epistemology, and everything is to be explained in terms of it. Knowledge first thinking shapes my own epistemological assumptions, so I will explain it in more detail in this introduction after my discussion of Interpretationism. However, this does not mean that I simply want to adopt Williamson's theory. I consider a range of test cases, whether of my own devising or adapted from the work of others, considering whether the theories of Williams and Williamson succeed in generating what I argue to be the most plausible interpretations of each case. I find flaws in both theories, and decide that the best course is to steer a middle ground between the two positions.

My own answer is that we should interpret agents by *optimizing their dispositions to know*. The habits of thought people follow can be or less suited to the acquisition of knowledge. I will tend to know more about the train schedule if I look at the stations timetable boards than if I listen in on the conversations of station staff. This is true even if listening in on the staff sometimes gains me knowledge, and the boards sometimes lead me astray. So between two alternative interpretations of me, one on which I believe what the boards say, and one on which I believe what I hear from staff, the one where I follow the boards is more likely to be correct. It is the interpretation on which I follow the best dispositions

[1] Williams 2020.
[2] Williamson 2007.

to know. This is only supposed to be an illustrative case to give a flavour of the view: in fact, it would be hard to spell the details out in such a way that this is an interesting interpretive dilemma. The advantage of my theory, I argue, is that it takes from Williamson a firm epistemological foundation that is lacking in Williams, and takes from Williams a sensitivity to dispositions that is lacking in Williamson. I further develop this view by applying it to some interesting edge cases, and by exploring whether it is at a disadvantage compared to rationality theories because the latter appeal to reasons for action whereas my theory considers only beliefs.

I: Interpretationism

Firstly, I am assuming that thought and talk is- at bottom, as it were - a matter of describing reality, and that proving a meaning for specific words and concepts is a matter of finding the specific chunks of reality to which they refer. This is certainly not a universal assumption. I will simply plead that this is surely *a* function of thought and talk: when I say ‘The cat is on the mat’ I am reporting that the world is a certain way. If the world is indeed that way, then what I say is true; if not, it is false. So much strikes me as obvious, and it is not now my concern to persuade anyone to who would deny it. I further think that is somewhere near ‘the bottom’ of what thought and talk do, because to my mind it is easier to understand how we can get from describing reality to other usages of language, such as jokes or metaphors, than vice versa. So I focus on thought and talk as descriptive of reality. Unfortunately, I do not have an account of exactly how exactly we can get to those other uses on my theory, however desirable that would be. This will present itself as an issue in the last chapter when discussing morality. All I can say is that such an account is required in the long run, but giving it is not within the scope of this thesis.

To say that language is a matter of describing reality is a picturesque way to say that meaning is *truth-conditional*. To give the meaning of ‘The cat is on the mat’ is to specify the conditions under which it is true: namely, the cat’s being on the mat. Besides being truth-conditional, meaning is also *compositional*. In English, there are finitely many recurring elements - the words - which can be combined to generate an infinity of sentences. When combined successfully, according to proper syntactic rules (and perhaps avoiding hidden failure conditions, arising, for instance, from problems of self-reference), each sentence will have a truth-condition, possibly one most speakers will never have contemplated before. These truth-conditions systematically depend on the words used and the way they are combined, so that competent speakers can understand unfamiliar sentences, such as ‘The pitiful helmsman hurriedly pardoned the contemptuous plumber’, simply through knowledge of the words and the rules for combining them. This is explained by the fact that each word makes a consistent contribution to the truth-conditions of the sentences in which it occurs. In whatever sentence the word ‘helmsman’ appears, the truth of the sentence depends on how things are with a specific set of the object, the helmsman. This set is the *extension* of the word helmsmen.[3] ‘Hurriedly’ will have to be given a more complex treatment along similar lines: a function, we might expect, taking us, in this instance, from the set of ordered pairs of *pardonings* to the set of ordered pairs of *hurried pardoning*s. Given the compositionality of language, we move from the truth-conditional content of sentences to the contents of specific words.

It is worth noting here that I have focused on language to explain this point, where it is most immediately obvious. I am, however, making the additional assumption, one that will prove relevant later on, that thought has a similar compositional structure. Complete thoughts present the world as being a certain way, just as sentences do, and are composed of recurring elements called

[3] Beyond the extension of the word, its *intension* also matters: an intension is a function from possible worlds to sets. Some think that words have *hyperintensional* contents that matter too: we will not wade into those debates here.

concepts. The basic justification for this is in the first instance natural analogy, and partly the direct connection between the two. I can hear the sentence ‘The pitiful helmsman hurriedly pardoned the contemptuous plumber’, learn that is meaningful in English and maybe that it is true. But something more has to happen for me to understand it, and part of that something more seems to be a matter of having a corresponding thought, that presents the world as being the same way as the sentence does, and composed of concepts corresponding to the sentence’s words, and sharing their content. If I have further thoughts about helmsmen, perhaps triggered by understanding sentences including ‘helmsman’ and perhaps not, it is natural to assume that the same helmsman concept is recurring between the distinct thoughts, just as the one word ‘helmsman’ recurs between distinct sentences. Though this is a controversial topic about which much could be said, having explained the point and its most basic motivation, I am simply assuming that thought is indeed compositional for the remainder of this thesis.

Although the point about thought is more controversial, what I have said about language so far has been fairly standard for analytic philosophy, the dominant tradition within Anglophone philosophy departments which has its roots in the work of Gottlob Frege: who, not coincidentally, did much to develop this approach to language.[4] This thesis is more specifically located within the Interpretationist tradition in the philosophy of language. Interpretationism approaches the question of meaning by taking the perspective of an interpreter, hence the name. We are interested in the theoretical question of what makes an agent’s words and thoughts mean what they do. This may be in spite of the fact that we are not at all puzzled by the practical question of what the agent actually means. Compare the earlier example about states: we may be perfectly confident that the United Kingdom is a state, but still want to know what makes it a state.

[4] My understanding of what analytic philosophy is owes much to Glock 2008. Frege 1997, edited by Michael Beaney, compiles much of his most influential work. There is also the counter-tradition that denies or complicates the truth-conditional picture that I have presented, prominently represented today by, for instance, Robert Brandom. See eg Brandom 1994.

Likewise, I may be confident about what my neighbour means by ‘dog’, but I still want to know what makes it the case that she means dogs. So I pretend that I do *not* know what she means. How would I go about discerning what she means? What would I do to interpret her? Answering this question, the Interpretationist holds, will help us answer the question that really interests us. The key to the theoretical question of why my neighbour’s use of ‘dog’ means dogs lies in entertaining the practical question of what her use of ‘dog’ means as if it were genuinely mysterious to us.

The reasoning behind this is fairly straightforward. Supposing that we are indeed able to interpret others correctly, then what we do during the process of interpretation must bear some connection to the reality of meaning. For instance, if we try to interpret others as believing truly, then our gaining knowledge by this strategy would be well explained by the fact that truth determines meaning. We can reverse-engineer the the way we determine the facts to discover, in David Lewis’s phrase, how the facts determine the facts[5]. Lewis himself drives the point home by insisting that we try to put ourselves in the perspective of an omniscient interpreter: at least, omniscient with respect to physical facts. This interpreter will determine the facts about meaning by surveying all of the physical facts and applying to them exactly the same principles by which the physical facts themselves determine the facts about meaning. The exercise is still illuminating despite our being very far from omniscient interpreters because we are already supposed to have a good grasp of the principles by which the physical facts determine the facts of meaning, which allows us to know a lot about meaning while being ignorant about a lot of physical facts. By idealizing away our extensive ignorance of physical facts, we bring the principles connecting the physical facts to the facts of meaning into sharper focus, since it is only knowledge of these principles that stands between the imagined interpreter and knowledge of meaning.

[5] Lewis 1974.

One general feature of Interpretationist approaches to meaning that is worth discussing is that they are holistic. We interpret agents by asking what makes best *overall* sense of them. While, as we have discussed, this will involve interpretations of specific words and concepts, these are not interpreted in isolation: meaning is not molecular. My neighbour's use of 'dog' does not refer to dogs simply because, for instance, it has the right sort of causal connection with dogs, but because the interpretation of my neighbour which is best overall is one on which it refers to dogs. In point of fact, the causal relationship between her use of 'dogs' and dogs may be one of the factors that makes this interpretation so good, but it can only ever be one of vastly many factors favouring the interpretation.

This raises a methodological issue. I often talk about the interpretation of specific concepts in isolation, as do other Interpretationists. How is this reasonable, given our holistic assumptions? Well, because the task of holistic interpretation is so complex, it is valuable to use idealized cases to focus on specific conflicts or principles. The isolation of words and concepts is incidental, what matters is the isolation of the principles, the cruxes around which our choice of a theory of interpretation may turn. This use of idealized cases is standard in many areas not only of philosophy but of inquiry more generally. Whenever I talk about choosing interpretations for a word or concept, it is to be understood that I am talking about choosing a holistic interpretation that involves the word or concept, and that there is an idealizing *ceterus paribus* assumption in place: for the purposes of the argument, the competing interpretations are assumed to be equal, except for the explicit differences over the word or concept in question.

Why, though, should we be holists in the first place? Because each individual word or concept is related to all of its peers within a complete system. Understanding our neighbour's use of 'dog' involves understanding the other words with which she combines it. Trying to interpret the word in isolation, accounting for non-verbal factors in the occasions on which she uses it, would give a radically incomplete picture. At the most basic level, it's crucial to know in each instance about its grammatical role in a sentence and the way it is modified, whether there is some kind of negation in force, etc. But there are further subtleties, such as discerning the precise extension of the term - that set of objects to which it refers - from plausible rivals. Does she mean *all dogs* by 'dog', or just the specific breeds of dog in connection with which she happens to have used the word? One of the factors that can help decide such questions is her use of such relevant terms as 'breed' and 'species'. It is easier to fix upon a unique solution by interpreting all three terms together than by stubbornly trying to interpret 'dog' on its own.

II: Knowledge-First Philosophy

So, when we attempt to interpret others holistically, upon what principles connecting the non-meaning facts to the facts of meaning do we, or would we, rely? In my first chapter, I survey the answers that have been given to this question in the literature. Cutting to the chase for purposes of this discussion, I find that the central recurring factor is epistemic success. We interpret others, as far as we can, as believing well. Thus, epistemology turns out to be crucial to meaning. Of course, epistemology is not the subject of this thesis, and I cannot spend too much time wading into questions of epistemology proper. Nonetheless, I should try to say something about my epistemological assumptions.

One of the later variants of Interpretationism which I discuss in more detail is Timothy Williamson's knowledge maximization proposal, which draws on his wider knowledge first philosophy. Since these ideas inform my own epistemological assumptions, I will explain them in some detail here. The basic idea of knowledge first philosophy is that knowledge should be placed at the centre of epistemology.[6] Knowledge enjoys explanatory priority - knowledge comes first - and other important phenomena flow from it, both within epistemology and beyond, as in the case of meaning which primarily interests us.

One of the starting points for this view is Williamon's harsh judgement on the attempts to define 'knowledge' in independent terms that have followed Edmund Gettier's famous argument that knowledge cannot be identified with justified true belief. The continuing proliferation of analyses, with each being continually adapted to accommodate new counter-examples, is diagnosed as a degenerating research programme. Theories become more complicated while doing less to increase our understanding of knowledge or epistemology more generally. Williamson's proposed explanation is that the concept of knowledge is primitive, and cannot be decomposed into constituent parts. If we accept the primitive concept of knowledge as our starting point, on the other hand, we can advance epistemology by using it to explain other things. For instance, he identifies evidence with knowledge, and argues that the norm of assertion is to assert only what one knows. He even reverses the usual order of explanation in analyses of knowledge by treating belief as a mental state that aims as knowledge. What he says about justification will be a significant topic later.

It does not follow from the fact that knowledge cannot be analyzed, however, that nothing worthwhile can be said about it. Aside from the idea already discussed that knowledge is the central phenomenon of epistemology, there are two important ways in which knowledge first philosophy characterises knowledge. Firstly, knowledge is the most general factive mental states. I can be in the condition of having stubbed my toe. I can be in the condition of being in pain. I can also be in the condition of being in pain because I have stubbed my

[6] Williamson 2000 is the central text of knowledge first philosophy.

toe. Of these, only the condition of being in pain is a mental state, a condition that is thoroughly a state of mind rather than anything else. Having stubbed my toe is a condition of my toe, while being in pain because I have stubbed my toe is a mental state combined with a further, non-mental condition. Of the mental states, some are propositional attitudes: stances taken towards ways the world might be, such as fearing that it will rain. A factive mental state is a mental state that you can only take towards truths: though you can fear that it will rain even if it doesn't, you can only see that it is raining if it is in fact raining. Likewise for hearing that it is raining, or remembering that it has rained. To say that knowledge is the most general factive mental state is to say that every one of these more specific factive mental states is a variety of knowing.

Another important claim about knowledge is that it follows the safety principle: a subject only knows that p if they could not easily have been mistaken whether p. Suppose that I am walking outside, and it is raining. I see and feel that it is raining: given that rain is falling all around me, the possibility of my being wrong about the fact that it is raining is remote. That is what knowledge is like. If, however, I form the true belief that it is raining because I see a screen displaying rain which I mistake for a window, that is another matter. Suppose the screen shows randomized weather conditions, so that it sometimes shows rain when it is raining, at other times an overcast sky when it is clear, a bright day when it is drizzling, etc. In this case, it could very easily have happened that I looked over at the screen and formed a mistaken belief about whether it was raining. My belief is unsafe, and so not knowledge. While Williamson maintains that safety is a necessary condition on knowledge, he does not think it is sufficient, or can be combined with further interesting conditions with which it is jointly sufficient. Hence knowledge remains primitive.

One final feature of knowledge first philosophy we should consider is its externalism. Four our purposes, epistemological internalism is the view that epistemic success, in whatever terms we want to think of it, entirely depends on factors internal to the subject. Assuming, for the sake of simplicity, some form of physicalism - the view that all facts depend upon the physical facts - then the obvious formulation of internalism is that epistemic success is entirely dependent on the internal physical state of the subject, with no relevant input from the subject's environment. Since part of the view that we are discussing is an externalism about mental states, it will not do to characterise an alternative to this physicalist formulation of internalism in terms of mental states. The most plausible formulation in this spirit, where we are concerned with strictly 'internal' mental states, is rather that epistemic success depends entirely on the subject's phenomenal states, their states of consciousness. This fits well with another strategy for explaining internalism, according to which an agent always has *access* to the determinants of their epistemic success: those determinants are present, or at least available, to their consciousness.

Epistemic externalism is the denial of internalism. Knowledge first philosophy is profoundly externalist. In the first instance, it treats knowledge as the key standard of epistemic success, and relates all other forms of epistemic success closely back to knowledge. Simply because knowledge is factive, this means that epistemic success is dependent upon the environment, the facts known by, and inependent of, the subject. More than this, however, knowledge is also characterised by a necessary safety condition, and safety too is an environmental factor. The weak relationship between the actual weather and the on screen weather which rendered my belief unsafe in the example above is no part of my internal condition, physical or phenomenal, and it is not accessible to me. This thoroughgoing externalism is particularly worth stressing, since it will tend to carry over to any theory of interpretation based on knowledge first philosophy. If the correct interpretation of an agent depends on what it would be epistemically successful for them to believe, and epistemic success depends on environmental factors, then the correct interpretation of an agent will also depend on environmental factors.

III: Thesis Structure

Before proceeding into the body of this thesis, I will give a chapter-by-chapter overview. The first chapter is a historical survey. I explore prominent figures in the Interpretationist tradition, to gain a sense of how Principles of Charity have developed over the 20th century. The three most important authors are clearly Willard van Ormand Quine, Donald Davidson, and David Lewis. They are my main focus, though for additional context I consider N.L. Wilson, an earlier author cited by Quine, as well Richard Grandy and Colin McGinn, who both contributed to the broader tradition. My main conclusion is that a plausible Principle of Charity should be framed in epistemic rather than alethic terms: we make best sense of a person by interpreting them as believing well rather than believing truly. Most authors accept some version of this idea, and even a philosopher as closely associated with truth as Davidson, and whose alethic tendencies, as we shall see, moved him to make some surprising claims, recognises the importance of the epistemic dimension of interpretation.

In the next two chapters, I examine two recently proposed epistemic Principles of Charity, starting with Robert Williams's work on rationality maximization in chapter 2. Williams takes Lewis as a starting point, and proposes that the correct interpretation of an agent's mental states is that which renders their actions most substantively rational in the light of their experiences. This theory is not completely neutral as regards epistemology: it assumes a notion of substantive rationality, and allows that there are some things that agents simply ought to believe in the light of certain experiences. However, it is very flexible, and someone sympathetic to the knowledge first epistemology described above might worry that rationality is too loose and internalist to constrain reference appropriately. I consider a range of cases designed to draw out this worry, including some adapted from Williamson himself. Given that the view is so flexible, however, it *can* avoid the problems raised here by plugging in an appropriate theory of rationality. Still, the theory as stated is flawed: a plausible account of rationality that can allay these concerns remains to be supplied.

I then consider Williamson's own theory of content in chapter 3, according to which the correct interpretation of an agent is that on which they know the most. Naturally, this theory is decidedly externalist, and easily handles the sort of case I raised to trouble rationality maximization. It is, however, vulnerable to other kinds of objection: we might worry that it is too simple, and so insensitive to the subtleties that something more like Williams's rationality theory can handle. I consider cases raised by M.G.F. Martin and Aidan McGlynn which suggest that knowledge maximization does not always deliver a decisive verdict where we would expect one. Williamson finesses his theory to handle Martin's case, and while this can be extended to cover McGlynn's, it enters the same sort of territory that Williams's did, where a fuller development of the theory is owed to deal with the issues arising. I press on with further cases of my own, designed to highlight the importance of dispositions to interpretation. After surveying the options available in the literature, I determine that what I want is a theory that combines the clear epistemological foundation of knowledge-based theories with more of the subtlety to be found in rationality-based theories.

This is what I try to supply in chapter 4. I mentioned above that what Williamson says about justification would prove important. In fact, he prefers the term 'rationality' for what accompanies true belief in Gettier cases. The relevant standard of rationality is that of conforming to good epistemic dispositions, dispositions which yield knowledge. A subject may sometimes follow dispositions to know without actually coming to knowledge, and this may occur even when the belief they form is true: this is what happens in classic Gettier cases. They have a true, rational belief that is not knowledge. What I propose is that instead of trying to maximize knowledge directly, we should instead interpret agents by maximizing this sense of rationality. The correct interpretation of an agent is that on which they are best disposed to know. This offers the firm epistemological foundation required to settle the cases that troubled Williams, and the subtlety required to settle the cases that troubled Williamson. It is the golden mean that I wanted at the end of the third chapter. I go on to explain why I think this theory is well-motivated as an application of knowledge first philosophy to the sphere of meaning.

This theory is developed for the remainder of the thesis. In chapter 5, I apply it to a range of interesting edge cases. First of all, I consider the staple sceptical scenario of a brain kept in a vat and stimulated to undergo experiences imitating those of ordinary human life. I argue that my theory delivers what is the most natural interpretation of such cases: an error theory on which they have plenty of of beliefs, many of which are errors about an external world. This is especially interesting inasmuch as I may outmaneuver not only more alethic versions of Interpretationism, such as that defended by Davidson, but also simple knowledge maximization. I next consider the case of the Swampman, a replica human who emerges fully-formed from a swamp. Again, I argue that my theory delivers the most plausible reading of this case: that the Swampman has beliefs much like ours. Finally, I consider a truly extreme scenario which amalgamates elements of both cases: the Vatbrain, a replica human brain that emerges fully formed into the setup of a classic BIV case. I do not believe that there is such an obvious interpretation here, but tentatively favour a proposal based on David Chalmers's notion of an extendible local matrix. The Vatbrain would thus have limited beliefs, of limited accuracy, about what is being directly simulated for them. This interpretation I consider to be plausible enough, and adequately supported by my theory.

In chapter 6, I consider an issue that has been postponed since chapter 3. Traditional rationality theories, such as those proposed by Lewis and Williams, take into account practical reason, or reasons for actions, and not just reasons for belief. Knowledge based theories, my own included, focus solely on belief. Does this leave my theory at a disadvantage? In general terms, I argue that it does not. Once we consider that we are interested in how all of the meaning-independent facts determine the facts of meaning, including those of which we are ignorant, some of the issues subside immediately. What may appear to be reliance on action in human interpretative practice can be seen instead as the use of actions as evidence for the facts about belief that do the real work in determining meaning. In more complex cases, where it may appear that avoiding some deep practical irrationality should trump avoiding a shallower theoretical irrationality in interpretation, the holistic and compositional nature of interpretation comes into play. The crucial trade-off is not shallow theoretical irrationality against deeper practical irrationality at an isolated point, but rather one shallow instance of theoretical irrationality against many instances of deep theoretical irrationality across the whole cognitive structure. Finally, I consider a broad area in which we might think that practical reason is especially important: moral terms and concepts. I argue that the crucial disposition in evaluating moral vocabulary is the disposition to employ it for moral enquiry. The reference of moral terms is just whatever it is we tend to learn about when engaged in moral enquiry. Given a realist moral metaphysics, this provides a nice clear answer to questions about the meaning of moral terms, and particularly to the worry that such meanings shift between different cultures. With different metaphysical commitments, the picture becomes murkier, though I argue that this is entirely appropriate. While I allow that various options could be combined with my broader theory, I tentatively suggest a mysterianism about moral meaning: in the absence of a realist moral metaphysics to anchor us, there are good reasons to think that we do not know enough to set about interpreting moral terms and concepts.

Chapter 1: The Development of the Principle of Charity

People mean things. They make utterances that mean things, such as ‘It’s raining’. They can also mean things without making utterances; they mean things ‘in their heads’ rather than ‘out loud’. How should we explain meaning? One popular way of answering this question appeals to the *Principle of Charity*. The following will discuss how this way of answering the question of meaning has developed over the late twentieth century. At a first pass, we may characterise principles of charity as principles governing interpretations, to the effect that a person means whatever it makes most sense for them to mean.[7] The core tradition of such principles begins with Quine and flows out to two key figures influenced by him at Harvard, Davidson and Lewis. Quine himself cites N.L. Wilson as a source[8], while Richard Grandy and Colin McGinn have approached the question of meaning in similar ways.

The plan for this survey is to explain the relevant context for each author, report what they say about the Principle of Charity, and address four questions, albeit not always in the same order: 1) how is the Principle motivated?; 2) to what is the Principle applied (eg to portions of language, language as a whole, language and thought both)? 2) is interpretation conceived as epistemic, with the Principle governing what makes the acts of interpreters justified, or is it conceived as metaphysical, with the Principle governing what makes a theory of meaning true or the phenomenon of meaning come to be?; is sense made epistemically, by attributing a person some favourable status such as knowledge, rationality, or justification, or is it made alethically, by attributing truths to her? To conclude each discussion, I will state the author’s principle as I interpret it. I

[7] ‘In short, folk psychology says that we make sense’ Lewis 1994: 1999 p 320.

[8] Quine 1960, p 59 (§13, footnote 2).

will largely refrain from evaluating principles of charity in what follows, except insofar as I deem such evaluations to serve an interpretive purpose.

I: Wilson

The ambition of Wilson's 'Substance Without Substrata' is to do away with bare particulars, 'the individual apart from its properties'.[9] What would the world be like, he asks, if Brutus had all Caesar's properties, and vice versa? Exactly as it is, he replies. Nonetheless, he notes that there would be distinct Carnapian state-descriptions for the actual and the property-swapped worlds, since, by Carnap's semantics, Caesar and Brutus are associated with different singular terms. What Wilson wants, therefore, is a semantics according to which the putatively distinct worlds answer to the very same complete description. To this end, he adverts to the Russellian expedient of replacing singular terms with definite descriptions.

This proposal he develops in two stages. At the first stage, he considers how a name comes by what we would (in his view, loosely speaking) call its reference. Suppose we want to know to what Charles' use of 'Caesar' refers. We know that he has used the term in five sentences. Though we understand fully the rest of his vocabulary, we are ignorant of Roman history. The sentences are as follows:

1. Caesar conquered Gaul.
2. Caesar crossed the Rubicon.[10]
3. Caesar was murdered on the Ides of March.
4. Caesar was addicted to the ablative absolute.
5. Caesar was married to Boudica.

[9] Wilson 1959, p 521.

[10] Incidentally, no one now knows what river is the referent of 'Rubicon' See Tom Holland's *Rubicon*. Aside from the general philosophical interest of the point, this rather suggests that the assumption of full understanding for the rest of the vocabulary was overly generous.

Wilson suggests that, to learn the reference of ‘Caesar’, we should conduct an (ever so ambitious) ‘empirical investigation, examining all the individuals in the universe’[11]. The individual we seek is that, if any, which: conquered Gaul, crossed the Rubicon, was murdered on the Ides of March, addicted to the ablative absolute, and married Boudica. As it happens, we discover one individual who was married to Boudica, and another who did or suffered the rest. Which should we posit as the referent of ‘Caesar’? Well, ‘we act on what might be called the Principle of Charity. We select as designatum that individual which will make the largest possible number of Charles' statements true’. So ‘Caesar’ turns out to refer to Julius Caesar.

Though this is of less concern to us, Wilson goes on to argue that we should not stop at reference, but treat ‘Caesar’ as a definite description: there is exactly one *x* such that most of F*x*, G*x,* R*x*…[12] Thence we are assured that a world in which Caesar and Brutus had purportedly exchanged properties is completely described in just the same terms as the actual world, and, on Wilson’s view, substrata may be eliminated from our metaphysics.

How does Wilson justify the Principle of Charity? The short sad answer is that he doesn’t, or anyway not directly. There are, however, two ‘default’ strategies we might apply on his behalf. The first emerges out of considerations of language as a practical tool. The meanings of words, as various of Wilson’s contemporaries would have been quick to point out,[13] are given by their use. Charles uses ‘Caesar’ to say that some man conquered Gaul, etc. So that is what

[11] Wilson 1959, p 531.

[12] This is getting a little ahead of ourselves, but we may note the conformity between what Wilson says and Lewis 1970. Caesar may be treated as a term introduced, and implicitly defined, by Charles’s Caesar Theory. Wilson discusses the Ramsey sentence of this theory, which, according to him, is the real semantic value of the conjunction of Charles’s sentences, as well, implicitly, as its Carnap sentence, which specifies what, loosely speaking, ‘Caesar’ refers to.

[13] Those of Wilson’s works that I can (quickly) trace appear to postdate Wittgenstein’s death, though he was certainly a contemporary of Austin and other Oxford ordinary language philosophers.

his use of Caesar means: a man who conquered Gaul, etc. Unfortunately for Charles, the man who conquered Gaul was not also married to Boudica, so there is no available meaning conforming exactly to his usage. It is at just this point, when we are considering which of several imperfect candidates to select as the referent of ‘Caesar’, that Wilson introduces the Principle of Charity. In this light, Charity is a matter of best fit with usage. Use gives meaning; Charles uses ‘Caesar’ in certain sentences; these sentences, we assume, are intended as reports about the past, rather than as fictions or jokes the truth of which is not at issue; so the meaning use supplies is whatever it takes to make as many as possible of Charles’s sentences true.

The second strategy looks at language in a light that would be unwelcome to the aforementioned contemporaries of Wilson, although we may suspect it differs in no important details from the first. Specifically, it emerges out of considerations of language as a vehicle for theory, and is suggested by the comparison to Lewis drawn in a footnote above. Charles holds a certain theory: a Theory of Caesar, or perhaps a Theory of the Fall of the Roman Republic. In order to state this theory, (it is pretended that) Charles adds a new term to his language: ‘Caesar’. The sentences of this theory function as implicit definitions for the new term. So far as interpretation goes, we may treat the assertion ‘Caesar conquered Gaul’ as if it were the stipulation ‘Let “Caesar” refer to whomever conquered Gaul’. Whatever object fulfills the most conditions given by these tacit stipulations is the referent of ‘Caesar’. The analogy between assertion and stipulation is the basis of the Principle of Charity.

What does Wilson take the Principle to apply to? To this question, at least, is given a clear, and restricted, answer: (superficially) singular terms. No wider role is considered. The presentation of the case, moreover, is unpromising for any such role: recall our assumption that the rest of Charles’s vocabulary is well understood. At a first pass, this assumption seems crucial to both mooted strategies for motivating charity. Unless ‘conquered Gaul’ is already interpreted, ‘Let “Caesar” refer to whomever conquered Gaul’ isn’t much of a stipulation. Likewise we could hardly say that Charles used ‘Caesar’ to speak of a man who conquered Gaul. As already noted, these two strategies are rather as one candle burning from both ends. Meanwhile, no attempt to discuss belief or thought apart from language is made. Wilson’s Principle applies only to the interpretation of singular terms in some language

The last two questions admit of similarly straightforward answers. Like later authors, Wilson dramatizes questions of interpretation by framing them as if we were seeking knowledge: in this case, of the reference of ‘Caesar’, as Charles uses it. But he is clear about where his real interests lie: ‘how’, he asks, ‘do words hook up with things?[14]’ To make progress with this question, he raises the sub-question: ‘how does a name in use get its significance?’. It is now that the epistemic turn is made, for Wilson says that this question ‘may best be attacked by asking another question : how should we set about discovering the significance which a person attaches to a given name?’[15]. He is dealing in epistemic means - the methods of discovering significance - to metaphysical ends - the basis of significance. Hence, of course, the obvious unreality of ‘examining all the individuals in the universe’. Interpretation, for Wilson, is ultimately a metaphysical matter.

Sense-making, meanwhile, is alethic. Wilson explicitly proposes ‘making the largest possible number of Charles’ statements true’. He does not stop to consider the epistemic status of Charles’ statements. One, we know, is false. For all we have been told, Charles could simply have memorized eight such sentences about Caesar in his youth, of which half were true and half false. He has simply been lucky in forgetting most of the falsehoods. Fanciful as this gloss may be, it is revealing enough that no attempt is made to exclude it. It is truth as such, not rationality or any of its kin, that Wilson’s Principle maximizes.

[14] Wilson 1959, p 528
[15] Ibid, p 529.

Summing up, Wilson's Principle of Charity is as follows*:* **the referent of 'a', where is 'a' is a singular term of some language *L*, is whatever it must be to maximise truth for the declarative sentences in which 'a' occurs that are affirmed by users of *L*.**

II: Quine

Imagine a linguist sent out to interpret the language of some remote tribe. Her goal is to produce a translation manual, with instructions specifying how any of the infinite sentences of the tribal language may be rendered into English. She has no knowledge of any related languages, and no interpreters to aid her. Such is the scenario of radical translation, as explored by Quine in *Word and Object*.[16] Quine has two core suggestions as to how the linguist should proceed. First, she should focus not on individual words, but on whole sentences. Second, her basic unit of analysis, at least at the beginning, should be what Quine calls the *stimulus meaning* of a sentence: more specifically, of an *occasion sentence.* By stimulation Quine has in mind something like sensory input, given a fairly rigorous scientific construal: he identifies visual stimulation with 'the pattern of chromatic irradiation of the eye'. An occasion sentence is a sentence assent to or dissent from which depends upon the stimulation which a subject has received. Quine's central example is the sentence 'Gavagai!', prompted by stimulations suggestive of the presence of a rabbit. The stimulus meaning of an observation sentence 'is a full cross-section of the subject's evolving dispositions to assent to or dissent from a sentence'.

However comfortable one might be with the primacy of sentences, those familiar with the traditional semantic concepts of truth, reference, extension, and

[16] Quine 1960 chapter 2, esp §7.

satisfaction are likely to find the world of radical translation, as Quine explores it, aptly alien. Things take a more homely turn in §13, which discusses the first translations of subsentential terms. Unsurprisingly, given the emphasis on whole sentences, the terms in question are logical connectives: the tools for building new sentences out of old. A term is to be translated as ‘not’ when its addition to a sentence uniformly turns assent into dissent; ‘and’ when it joins two affirmed sentences into one affirmed sentence, but the addition of any disavowed sentence results in disavowal; and so on. What recommends this strategy? ‘One’s interlocutor’s silliness, beyond a certain point, is less likely than bad translation’.[17] In a footnote, he compares this point to Wilson’s Principle of Charity.

What is Quine’s justification for invoking Charity? It appears to rest on the probability of errors, and in particular the relative probability of certain errors. A helpful reference point may be Hume’s argument against miracles.[18] However weighty the testimony alleging a given miracle, Hume contends, the falsity of the testimony is always more likely than the reality of the miracle. Quine, like Hume, is offering a choice between two unlikely options, and urging us, or rather the radical translator, to take the least unlikely. In Quine’s case, the options are logical error on the part of the tribe on the one hand, or translation error on the part of the linguist on the other. Quine thinks that consistent logical error throughout the whole tribe is extremely unlikely. Such errors are beyond the point of tolerable silliness. Radical translation being a difficult task, an error on the linguist’s part in trying to make sense of the putative connective-terms is rather more likely. Thus the linguist should interpret the tribe as logically competent as far as she can, and try to isolate terms eliciting patterns of assent and dissent much as the English connectives do, which are then translated using those very English connectives.

[17] Ibid, p 60.

[18] Hume 2007 [1748], Section X, ‘Of Miracles’.

What Quine takes the scope of charity to be is a delicate matter. Certainly he makes explicit appeal to it only in discussing the logical connectives, but he does not appear to restrict its application to the connectives. Later, Quine discusses the translation of non-logical vocabulary. In keeping with his sentential focus, he prefers not to speak of terms translating terms, but of ‘analytical hypotheses’ adduced to aid the real business of translation, from sentence to sentence. Here he does not exactly advocate the Principle of Charity as a method, but he does stress the desirability of avoiding the attribution of ‘absurd or exotic’ beliefs: ‘for translation theory, banal messages are the breath of life’.[19] This seems quite right. Quine himself points out that logic error is ‘only the extreme’ of a continuum of possible errors. There are many kinds of error we might think less likely than errors in our translations: errors, for instance, about whether food is poison, night is day, or joyously dancing children are dead. By the reasoning explained above, the linguist should eschew any analytical hypotheses leading to the attribution of such errors, unless doing so ‘would seem to call for much more complicated analytical hypotheses’.[20] Thus Quine’s version of Charity applies to logical vocabulary foremost, but also, if weakly, to whole languages.

Let us deal first with the question of sense-making before moving on to the larger one of interpretation. Despite quoting Wilson’s explicitly alethic formulation of Charity, it seems that Quine sees matters more epistemically. His interest is not so much in what is merely false, but in what is startling, silly, exotic, or absurd. This is also clear from the way we have seen Charity motivated: some errors among those translated may be less likely than errors in the translation, but it is not natural to suppose that all are. Doubtless the tribe would struggle to get its cosmological ideas into any reputable physics journal. Strictly speaking, however, probability is all, and the invocation of any explicitly epistemic category would be out of keeping with the general tenor of Quine’s approach. We may well suspect that probability of error bears some significant

[19] Quine 1960 p 69.
[20] Ibid.

relationship to rationality of error, but positing such a relationship for exploitation in our translational practice goes beyond what Quine himself endorses. Nonetheless, Quine’s Principle is closer to being straightforwardly epistemic than it is to being straightforwardly alethic.

A further peculiarity of Quine’s Principle is that it is minimizing rather than maximizing. Quine counsels the avoidance of extreme error-attribution, rather than the embrace of some positive strategy. The real role of Charity for Quine is as a check on translational work already conducted by other means. Both this and the failure of his Principle to be either straightforwardly epistemic or straightforwardly alethic is explained by a more general peculiarity of Quine’s approach to translation already noted: his emphasis on the apparatus of stimulus meaning in preference to truth. Indeed, strictly speaking, Quine would not want us to think of minimizing error at all, but as indirectly conforming the stimulus meaning of proposed translation sentences to the stimulus meaning of the tribal sentences that they are supposed to translate. Stimulus meaning admits of more direct empirical investigation than either truth-value or epistemic status, and so Quine wants it to do more theoretical work than either.

This last point brings us directly to our final question: does Quine think of interpretation as metaphysical or epistemic? The distinction is at least a little artificial, since each philosopher is usually interested both in the basis of meaning and our knowledge of it, but Quine is strongly skewed towards the epistemic. In *Word and Object*, Quine has broad epistemological ambitions: explaining how language is a fitting vehicle for scientific theory. This is how I understand the drift of Chapter I, and the point of pondering ‘our talk of physical phenomena as a physical phenomenon, and our scientific imaginings as activities in the world that we imagine’.[21] The object is to show how even the sentences of the most remote and austere theory relate back to our sensory stimulations, the ultimate evidential basis of all science. This cannot but be done by proposing

[21] Ibid, p 5.

more austere theory, in this case, theory of language, and consequently Quine is at pains to explain the empirical respectability of this theory. By programmatically indicating how one might build a full theory of an alien language on the most rigorous empirical basis, he is offering a promissory note for his programmatic account of language in general. It is in this context that Quine proposes his account of radical translation, including its limited role for the Principle of Charity.

This, then, is Quine's Principle of Charity: **the translation of some sentence *S* of some object-language *L*, to which the users of *L* generally assent, by some sentence *S1* in some meta-language *L1*, from which the users of *L1* generally dissent, even given relevantly similar histories of stimulation, is to be avoided, and in difficult cases careful checks should be made that no superior translation may be produced.**

III: Davidson

Karl says 'Es regnet'. Some of us know that he has said that it is raining. Davidson wants to know how such knowledge is possible.[22] Firstly, he wants to know by means of what theory could we predict the meaning of this particular utterances of Karl's: one of an infinity of utterances he might have made and whose meanings his hearers might have known. Secondly, he wants to know by means of what evidence could such a theory be supported.

Davidson's answer to the first question is that we could know a theory of truth for the language Karl speaks. This would a theory along Tarskian lines,

[22] Davidson 1973.

predicting, for every sentence of the language, a sentence of the form '"P" is true in language L if and only if P'. This is done by taking the concept of satisfaction, which relates sentences to sequences of objects, as primitive. The axioms of the theory state the conditions under which the simplest open sentences are satisfied, and how the satisfaction of how complex sentences depends on the satisfaction of simple sentences. Truth for all closed sentences is defined in terms of satisfaction. Given some means of mapping sentences that do not appear amenable to such treatment (eg, those featuring indexicals) on to more complex sentences that are (sentences which might be said to exhibit the logical form of the simpler but recalcitrant sentences), an entire language can be thus interpreted.

The next matter to settle is just what evidence might support such a theory. Davidson is only interested in evidence available to a theorist who does not already understand the language in question: this is a project of *radical interpretation*, as Quine's is of radical translation. His suggestion is that we may identify which sentences Karl takes to be true. This, admittedly requires knowledge of the psychology of the interpreted agent, especially how they manifest the attitude of holding-true.[23] The core assumption is that utterances tend to be sincere assertions. In fact, we know that this is severe idealisation, as far as human utterances go, but the most effective procedure is likely to involve provisionally identifying utterances with sentences that are held true. Additional evidence may then be sought as confirmation or disconfirmation: if those listening to the speaker laugh, this might be taken as evidence that the sentence is not held true; if the hearers respond aggressively, and the speaker repeats themselves aggressively, this might be taken as evidence that the sentence is held true by the speaker, if not the hearers.

Once we have begun to identify sentences held true, we may then consider whether some sentences are held variously true or false in different

[23] Which raises the question, pressed, as we shall see, by McGinn, of *more* radical interpretation. Might there be creatures who, for instance, communicate with one another solely through what we would call riddles?

circumstances. Suppose we observe that Kurt only utters ‘Es regnet’ when it is raining. Sometimes others (who, say, stand at some distance from a window down which condensation is running on a cloudy day) say ‘Es regnet’ and Kurt (who is immediately beside the window) appears to challenge them. This provides evidence that Kurt holds ‘Es regnet’ to be true if and only if it is raining; and so that, in Kurt’s language, ‘Es regnet’ is true if and only if it is raining.

To review that last inferential step. ‘Kurt holds ‘Es regnet’ to be true if and only if it is raining’ is somewhat ambiguous. The point is not to attribute Kurt any beliefs about the truth conditions of ‘Es regnet’: according to Davidson, we must discover the truth conditions of sentences for ourselves before we attribute any beliefs, regarding truth-conditions or otherwise, to those we interpret. It is merely to state a regularity in Kurt’s holding true of ‘Es regnet’: if it’s raining, he holds the sentence true, but if it’s not, he doesn’t.[24] So what exactly is it that justifies our inferring, from this regularity, that ‘Es regnet’ is true if and only if it is raining? The answer, of course, is Charity: we are ‘to maximize agreement, in the sense of making Kurt (and others) right, as far as we can tell, as often as we possibly can’.[25]

In contrast to some others, Davidson is explicit about the motivation for Charity. In fact, there are two distinct motivations Davidson gives, though one is very much dominant, and he does not appear to have distinguished them himself. There is *Charity as a general prerequisite for interpretation*, appearing in

[24] Are there not many such regularities? If he holds the sentence true, then it’s raining and this hour will last 60 minutes; then raindrops are striking the ground; then there is fast, predominantly liquid precipitation, etc. Davidson 1976 addresses such questions, concluding that understanding a language involves not merely knowing a T-theory for the language, but knowing of that theory *that* it is a T-theory for the language. Some regularities, being not just true but entailed by an adequate theory, are really counter-factually robust laws. Such a response, however, avails little when one is only beginning to construct a T-theory. One possible answer is that this does not matter: any regularity will do to get the business of theory-building under way, and as more evidence is collected the theory will be more refined. Another possibility is to employ the knowledge of the agent’s psychology assumed earlier: some conditions, such as rain, might be thought more salient to the agent than others, and so regularities involving those conditions more important.

[25] Ibid: 1984 (2001) p 136.

passing in ‘Radical Interpretation’, and *Charity as a specific prerequisite for interpretation*, also suggested there but discussed more fully in later papers. What I mean by Charity as a *general* prerequisite for interpretation is the idea that, unless we attribute some thing with a tendency to be right rather than wrong, we will have no interest in treating that thing as a subject of interpretation. The ability to distinguish truth from falsehood is the mark of a rational being. This is, of course, a *mere* ability, flawed and finite, rather than a superpower, but absent any such ability we will have no reason to treat something ‘as having beliefs, or saying anything’ at all.[26]

Charity as a *specific* prerequisite for interpretation is the idea that, for any given sentence a speaker holds true, the more true beliefs we attribute to that speaker, the better placed we will be to understand the sentence in question. Suppose someone says ‘There is milk in the fridge’. They be speaking truly or falsely. But unless I can attribute them with the belief that there is a fridge in the house; that fridges are cold; that it matters whether milk stays cold; then I will have little reason to think that it is a *fridge* and *milk* that they are talking about. ‘The more things a believer is right about’, Davidson says, ‘the sharper their errors are’. Charity is needed because without the illumination of true beliefs in the background, we cannot discern the sense of any sentence we may bring into the interpretive foreground.

Davidson thinks that Charity applies to both belief and language, though his main focus is language, and contends that belief is dependent on language. First, it is clear that Davidson is interested in interpreting whole languages: the point of invoking a Tarski-style truth theory is that ‘it entails, for every sentence *s* of the object language, a sentence of the form: ‘*s* is true (in the object language) if and only if p.”[27] The Principle of Charity is then only explained after Davidson says explicitly that the last step is to interpret ‘the remaining sentences’: ie, all those

[26] Ibid p 137.
[27] Ibid p 130.

sentences, unlike ‘Es regnet’, ‘whose held truth value does not systematically depend upon the environment’. Charity is a general method for producing a complete truth theory for the language.

Belief enters the picture only once we are forced to distinguish between what is held true and what, in the relevant language, is in fact true: that is, when, reaching the limit of Charity, we attribute error.[28] This, presumably, is what we shall have to do with the speaker encountered earlier, whose utterance of ‘Es regnet’ at some distance from a window on a cloudy day Kurt challenged. Once the concept of belief has been introduced, however, we may count a new belief for every sentence that is held true. We even attribute beliefs without associating them with any sentence a given speaker has uttered or heard, as when we assume our interlocutor believes that fridges are cold when she says that there is milk in the fridge. If we can identify some sentence of the interlocutor’s language of which this belief is a holding true, all the better, but this is inessential to the immediate interpretive task. Hence Charity does apply to beliefs: Davidson would have us make out, as far as possible, that what Kurt holds true really is true; and, for every every sentence the holding true of which is explicitly manifest in Kurt’s linguistic behaviour, presuppose a background of true beliefs behind it. Nonetheless, the use of ‘background’ here is telling: it is always sentences, rather than beliefs, that are in the foreground for Davidson.

Davidson is consistent in preferring a mixed account of sense-making. Already in 1973, Davidson is sensitive to ‘intelligible error’ and ‘the relative likelihood of various kinds of mistake’.[29] Thus, within the bounds of that paper, he moves from talk of *maximizing* agreement to talk of *optimizing* it[30]. In later work, cognizant of others’ criticism, he is still clearer on this point. Nonetheless, that criticism was not for nothing, and Davidson’s bias is towards using truth as the

[28] Davidson 1975, especially p 170.
[29] Davidson 1973: 2001 p 136,
[30] Ibid p 137.

primary instrument of sense-making, albeit taking certain precautions, rather than using an epistemic category in its place.

This comes out best in the discussion of Charity in ‘Thought and Talk’, where Davidson airs the rather surprising view that no one has ever really believed that the earth is flat.[31] The basic point is that which I have made less dramatically with the sentence ‘There is milk in the fridge’: past a certain point of error, it becomes doubtful whether someone is really speaking of the fridge, or the earth, at all. To take seriously the possibility that the hypothesis of a flat earth passes this point of error, however, betrays a radical reliance on truth as the basis of sense-making. Anyone inclined to a more epistemic view would want to pay more attention to the arguments historically given against the hypothesis: until sufficiently compelling evidence was available, such as the observations of different stars from sufficiently distant locations, we would have no reason to regard the flat earth hypothesis as suffering an especially unfavourable epistemic status, and thus no reason to doubt that people believed that the earth is flat.[32] Davidson is happy to ignore such points precisely because truth is what matters to him, even if his considered position is that truth need not be maximized come what may.

Davidson has similarly mixed views on what interpretation amounts to. He begins ‘Radical Interpretation’, as I have intimated, by asking ‘what could we know that would enable us to’ interpret Kurt’s utterance of ‘Es regnet’, and ‘how could we come to know it?’[33] Yet he begins his introduction to *Inquiries into Truth and Interpretation* by asking a different question: ‘What is it for words to mean what they do?’[34] Clearly, then, Davidson is interested in both the epistemology and the metaphysics of meaning. Nonetheless, the main focus in

[31] Davidson 1975: 2001 p 168.

[32] See Aristotle’s *On the Heavens*, II.14, 297-8, for discussion. Refuting the No Flat Earth Hypothesis hypothesis is not really my concern, but I must add that it is surely more charitable to Aristotle to accept that he was arguing against people (he cites Anaximenes, Anaxagoras, and Democritus) who really did believe that the earth is flat.

[33] Davidson 1973: 2001 p 125.

[34] Davidson 1984 (2001) p xv

his papers is on questions of knowledge. Unlike Lewis, he does not drop any hints that the epistemic framework is not to be taken seriously, but dwells earnestly and often on what the evidence for a theory of meaning should be.

This is because Davidson, unlike Lewis, is not a dogmatist. That is, the question of how knowledge is possible, both in general and in specific domains, is a problem for Davidson as it is not for Lewis. So much is clear not only from the epistemic focus of his early work on radical interpretation, but the later use to which that work is put. In 'A Coherence Theory of Truth and Knowledge', Davidson attempts to leverage his views on interpretation into a transcendental argument against scepticism.[35]

The details of that attempt do not now concern us; the point is what it reveals about Davidson's philosophical concerns. It matters to Davidson not merely what knowledge is, but whether and how we come to know. The Principle of Charity is central to Davidson's answer to these questions: both in their general form, and as regarding the specific domain of meaning.

Davidson's core Principle of Charity may be stated thus: **that some speaker holds a sentence S true in and only in condition C is evidence that S is true if and only if C obtains.**

IV: Grandy

The goal of 'Reference, Meaning, and Belief' is to refine Quine's work in *Word and Object* to provide a philosophical account of translation.[36] Like Quine, and the broader radical interpretation tradition which he inaugurated, Grady's point is to illuminate more general issues of linguistic interpretation that apply even

[35] Davidson 1983.
[36] Grandy 1973.

between speakers of the same language[37]. He holds that the main purpose of translation is to predict behaviour. Once we translate utterances, we may infer desires and beliefs; once we infer desires and beliefs, we may predict behaviour. Ideally, we would know our subject's beliefs and desires perfectly, and so be in the best possible position to predict their behaviour. But the means by which we acquire such knowledge is the very issue at stake. This would seem to leave us at an impasse, but Grandy suggests an escape route. We do, at least, know about the structure of *our* beliefs and desires. Using that structure as a model, we can predict our subject's behaviour by considering what we would say and think and do in their position. Hence Grandy offers the Principle of Humanity as a constraint of translation: interpret others by analogy with ourselves.

Primarily interested in translation though he may be, Grandy is explicit, as we have seen, that translation and linguistic interpretation more generally are just one element in a wider epistemic project: predicting behaviour. The Principle of Humanity applies, therefore, not only to language but also to beliefs and desires. We have also seen what Grandy's motivation is: our knowledge of our own attitudes is the most pertinent evidence we have available to us when interpreting others.

Further, the centrality of prediction on Grandy's view means that even once we have made translations and attributed attitudes accordingly, there is still work to do. 'If a translation tells us that the other person's beliefs and desires are connected in a way that is too bizarre for us to make sense of', he says, 'then the translation is use-less for our purposes'. So Humanity becomes 'a pragmatic constraint' on translation.[38] I take it Grandy has in mind something like the following. Suppose we translate some utterances of our subject as implying that he desires to dance with his bride, and believes that he will look clumsy if he does so. Modelling the subject on ourselves, we will assume that, while he does

[37] Unlike Quine and others who followed him, Grandy is not interested in *radical* translation. This is important. Contrast McGinn's concern with *more* radical translation.

[38] Grandy 1973 p 443.

not want to look clumsy, his desire to dance with his bride, bound up as it is with eg strong social expectations about a crucial rite of passage, will prove overwhelming, and so predict that he will indeed dance with his bride. Suppose instead we translate his utterances so as to imply that he desires to dance with an abacus, and believes that he will halve Leeds United's shots on goal if he does so. These attitudes are surely so mystifying that we will simply be unable to make any useful predictions about the subject's actions. Does he want Leeds United to succeed? Does he think that, if their shots on goal are halved, those shots will find the back of the net fifty times more often? All bets are off. For our translations to serve their purpose, we have no choice but to model others on ourselves.

As to whether Grandy regards sense-making as alethic or epistemic, the answer again should be clear: it is neither. Instead, he thinks of sense-making as analogical: we make sense of others by analogy with ourselves. If we would believe falsely or unreasonably in another's shoes, we should attribute false or unreasonable beliefs to that other agent.[39] Nonetheless, it is clear that this analogical view is closer to an epistemic than an alethic alternative. Grandy argues from cases for the superiority of his own view over a truth-maximizing view.[40] Paul arrives at a party and says 'The man with a martini is a philosopher'[41]. In fact, he is looking at a man who, though drinking a martini, is not a philosopher; while in the garden out of sight is a philosopher drinking a martini. If we are maximizing truth, we should interpret 'The man with a martini' as referring to the man in the garden, for then the whole sentence is true. Yet surely that verdict is mistaken: Paul is talking about the man in front of him, whom he mistakes for a philosopher, and not another man of whose presence he was not previously aware. This, says Grandy, is just what the Principle of Humanity predicts: in Paul's place, we would believe nothing and say nothing

[39] Compare Lewis 1974.

[40] He attributes this view to Quine, in my view dubiously, and oddly, given that he had earlier discussed the importance of the obvious for Quine.

[41] Grandy 1973 p 445.

about the philosopher in the garden, while we might easily form false beliefs about people who are before our eyes (because of a resemblance, say, or a misread visual cue).

He then proceeds to add some epistemic content to his view. Why would we not attribute Paul with any beliefs about the philosopher in the garden? Because, Grandy says, we hold a causal theory of belief. Paul had no causal connection to the philosopher in the garden, and so he could not have had beliefs about him. From this causal theory of belief, Grandy derives a causal theory of knowledge, given that knowledge requires belief; and given his theory of translation, he derives a causal theory of reference. Grandy's account of sense-making is fundamentally analogically, but it is closer to a epistemic account than an alethic one.

Finally, we should consider whether Grandy's account of interpretation is epistemic or metaphysical. As usual, the answer is mixed. For the most part, however, his focus seems to be on how we can and should go about our business of translation. The following is a telling comment: asking about what evidence may be employed for the translation of foreign languages gives us 'some idea of what it means to say that one translation is preferable to another. Depending on one's view of the connection between evidence and truth, it also says more or less about what it is for a translation to be correct'. Primarily, Grandy is interested in how we come to know about meaning; secondarily, he suggests that what he says here is relevant to questions about what meaning itself amounts to, but is happy to let the reader decide just how relevant. Nonetheless, his later willingness to derive a causal theory of reference from his views on translation suggest that he considers it to be very relevant indeed.

Grandy's Principle of Humanity may be stated thus: **we should translate others in whatever way best conforms to the hypothesis that they believe and desire as we do.**

V: McGinn

In ‘Radical Interpretation and Epistemology’, Colin McGinn tries to amend Davidson’s account of radical interpretation to avoid the sweeping anti-sceptical consequences elaborated in ‘A Coherence Theory of Truth and Knowledge’.[42] In the first instance, he does so because he is incredulous of those consequences: not so much because he is himself a sceptic, as because he does not imagine that the sceptic can be so easily answered. There is further motivation, however: McGinn explicitly pushes the boundaries of radical interpretation into *more* radical regions.[43] Although he discusses more outlandish cases yet, his main focus is on how we should set about interpreting a brain in a vat.

McGinn presents Davidson as interpreting a BIV as follows.[44] The BIV ‘says’ (presumably it is hooked up to a device that converts neural activity into speech) ‘There is a round red thing before me’. Thus, Davidson concludes that it takes ‘There is a round red thing before me’ to be true. But the condition of there being a round red thing before it does not obtain; indeed, it never does when the BIV manifestly takes that sentence to be true. But other conditions do coincide with these takings true: the condition, say, of an electrode ‘sending *n* volts into their occipital lobe’. But according to Davidson’s interpretive strategy, we are quite generally to interpret a sentence as being true in just those conditions in which it is held true. So the meaning of ‘There is a round red thing before me’ is given by the following biconditional: ‘There is a round red thing before me’ as uttered by the BIV is true if and only if an electrode is sending *n* volts into the BIV’s occipital lobe. Thus even the BIV is preserved from massive error.

[42] McGinn 1986.
[43] Ibid p 190
[44] Ibid p 186

McGinn finds this interpretation implausible. Apart from the conviction that any envatted brain *must* be in massive error (save, perhaps, in the case where whatever machinery is acting upon it is specially calibrated to preserve it from error), McGinn argues from considerations in the philosophy of mind. The BIV, we may suppose, has an un-envatted twin. The brain states of the BIV and the twin are identical. Whatever neurons are firing in the one fire also in the other.[45] But, according to Davidson, their mental content is radically different. McGinn is prepared to accept that content seeps out of the head at least to some extent, but not that the state of the brain is as irrelevant to the contents of belief as Davidson's approach to interpretation would suggest in this case.

McGinn's alternative proposal is as follows. He would have us distinguish two kinds of content: belief in general, and experience in particular. With sufficient knowledge of a subject's perceptual system, we can ascribe a course of experience to them. This experience, we assume, is sufficient to yield a core of observational concepts, such as those of redness, roundness, and the relation of being before. But experience and observational concepts are not enough to yield belief: we must determine what beliefs the experience induces. One may undergo an experience as of a red round object, but remain resolute in the belief that one is a brain in vat: or, less outlandishly, that one is dreaming. McGinn's answer is that we should interpret subjects as tending to trust their experience. Accordingly, he interprets the BIV who says 'There is a red round object in front of me' as saying, falsely, that there is a red round object in front of it.

McGinn's departure from Davidson has already been explained. His justification for the principle that we should interpret others as trusting their experience is simply that we cannot avoid doing so: such trust is '*a condition of interpretability*'.[46] What McGinn wants from an interpretation, meanwhile, is 'a total set of psychological and semantic ascription'. McGinn has both language

[45] Ibid p 186.

[46] Ibid p 193. Cf Grandy's justification of his analogical procedure.

and attitudes in view. His account of sense-making is radically non-alethic, explicitly allowing as it does the possibility of massive error. It is instead an epistemic sense-making tied to a firmly internalist epistemology, in the broad tradition of phenomenal conservatism. Finally, McGinn says enough to indicate that his account of interpretation is really epistemic rather than metaphysical. Upon considering the case of a being that systematically distrusts its own experience, he follows the justification of his Principle of Charity with the observation that ‘this is not to say that such a person is *impossible*; it is just that he is not *interpretable*’. Clearly, then McGinn cannot take his Principle of Charity to govern directly the facts of meaning and belief, else this space between the interpretable and the possible would close.

McGinn’s Principle of Charity may be stated thus: **a subject S is to be interpreted by A) attributing a course of experience to S; B) assuming that S generally trusts their experience; and C) attributing contents to S’s utterances and attitudes that conform with the foregoing attribution and assumption.**

VII: Lewis

Take any person: Karl, for instance. Psychologically, he is characterized by the attitudes that he holds: his beliefs and desires. He also uses a certain language, such that when he utters an indicative sentence he describes the world as being one way rather than another. What makes it the case that Karl holds the beliefs and the desires that he does, rather than any others? What makes it the case that he uses the language that he does, rather than any other?

Lewis's answer is that Karl believes and desires what is rational and what rationalizes.[47] Karl's beliefs are rational in that they are the result of some rational inductive system.[48] Karl takes in perceptual evidence and rationally forms new beliefs in response: that there is a rabbit, on seeing a rabbit. He also makes rational adjustments to his prior beliefs: he abandons his assumption that he is not in rabbit-country. This latter is not merely a question of updating appropriately according to new evidence, but of having had a rational assessment of the probability of the newly accepted hypothesis all along. Karl never imagined himself to be so far away from rabbit-country that he immediately concluded, on seeing this rabbit, that it must have dropped from a plane. Karl is even rational in his basic desires: no mugs of mud, but coffee will do just fine.

The results so far have been pretty vague: Karl is left rationally believing and desiring somehow. The main factor fixing his attitudes more definitely is the course of his perceptual experience. But that is not the only factor, for rationalization comes into play. Karl does not merely perceive, believe, and desire: he also acts. When he does, he does so rationally: that is, he acts in ways that he believes will satisfy his desires. If he walks to a café and orders a panini, then he desires a panini and believes that he can acquire one in the café he visits. The course of Karl's perception and the detailed history of his behaviour, along with background constraints of rationality, jointly determine his attitudes.

One important dimension of Karl's behaviour is his linguistic behaviour. Quite apart from its role in fixing Karl's attitudes, it is to this that we must turn to address our second question, that of the determination of Karl's language. Lewis's proposed mechanism is as follows: Karl linguistic behaviour conforms to certain conventions of truthfulness and trust shared by his wider linguistic

[47] For the most part I am following Lewis 1994 p 320.

[48] The reference to an inductive system harks back to Lewis 1974, though I have subtracted the subjectivity, in keeping with the more objective tenor of Lewis 1994.

community.[49] Karl generally utters a sentence only if he believes it to be true; on hearing a sentence, meanwhile, he generally takes it to be true. But truth is relative to a language: Karl is not concerned about uttering truths in Klingon. Karl's language, therefore, is just that language in which he is truthful and trusting[50].

Because Lewis's views are worked out in such detail across so many writings, the basic answers to my four questions are evident enough in merely presenting those views here. There are still further issues to address. To begin with one clear point: Lewis applies the Principle of Charity primarily to attitudes, and secondarily to language, through the mediation of linguistic conventions as discussed above. A second clear point is that, at least as regards its primary application, Charity is thoroughly epistemic, concerned with maximizing rationality in a substantive sense: even basic desires and prior probabilities are rational. There are, nonetheless, some interpretive difficulties regarding the latter point that the former point enables us to resolve.

In some cases, epistemic Charity is conspicuous by its absence. In 'Putnam's Paradox', Lewis rebuts Putnam's model-theoretic argument for the radical underdetermination of reference.[51] To do this, he invokes his famous naturalness constraint: some objects and properties, specifically those distinguished by their objective similarity[52], are more eligible referents than others. The class of rabbits is a fairly, if not perfectly, natural property, but the extension of 'is a rabbit' under some tricksy permutation is highly unnatural. The point that bears notice is

[49] Lewis 1974, 1975. For this purpose, a convention may be characterised as a regularity sustained by common interest and knowledge. Fuller definitions are given in both papers, but they are long, and nothing here turns on the detail.

[50] Though Lewis himself does not explore the idea, attributing Karl with conformity to the relevant conventions is presumably an instance of rationalization: Karl utters the sentences that he does because he believes them to be true, and desires to communicate truths. We can further rationalize his linguistic behaviour, and thus uncover further determinants of his language, by attributing him with conformity to standard pragmatic maxims, concerning eg informativeness.

[51] Lewis 1984.

[52] An object is similar in that its parts are similar: a ball-bearing is internally similar, and so natural, while the sum of a stick and some arbitrary portion of the atmosphere of Venus is not. So too for a property, except properties have elements rather than parts, *pace* Lewis 1991.

that Lewis here simply invokes his naturalness constraint without offering any motivation. This is curious, since in the earlier 'New Work for a Theory of Universals', Lewis presents the constraint as arising (somewhat mysteriously) out of the Principle of Charity.[53]

Another case is 'Psychophysical and Theoretical Identifications'. Here Lewis considers the reference of various psychological terms such as pain, as well as (we will be returning to this later) belief and desire. In line with 'How To Define Theoretical Terms', Lewis proposes that these are terms 'introduced' by the theory of folk-psychology, and that the 'platitudes' of this theory implicitly define the terms. The question is whether Lewis sets any store by the epistemic status of these platitudes. At first sight, he does, for he requires the platitudes to be 'common knowledge'. Yet he goes on to say that the sentence which provides implicit definitions for our psychological terms is not one conjunction of all these platitudes, but rather the disjunction of all conjunctions of most of them: for 'that way it will not matter if a few are wrong'.[54] Evidently, then, he is not requiring that we *know* the platitudes of folk-psychology, and the common knowledge he speaks of is best understood as general *familiarity* with the platitudes. Again, no fully epistemic Charity is in sight.

All this, however, is readily dealt with when one considers that Lewis applies fully epistemic Charity only to attitudes, and in neither case are attitudes primarily at issue. In 1984, Lewis chooses to 'acquiesce in Putnam's linguistic turn', stating explicitly that in 1983 he had 'relocated' Putnam's problem to the sphere of attitudes.[55] In 1972, Lewis is discussing the meaning of psychological terms within public language. So long as we all intend to speak truly when we state folk-psychological platitudes, that guarantees that our psychological terms refer to whatever they must in order for the platitudes to be true. Lewis's theory

[53] Lewis 1983.
[54] Lewis 1972 p 258.
[55] Lewis 1984 pp 57-58.

makes no additional demand in the form of epistemic constraints applying directly to language.

Lewis's treatment of psychological terms is further notable, since it is whence his motivation for the Principle of Charity derives. According to the theory of folk-psychology, we have beliefs and desires, and these beliefs and desires are rational. So whatever it is that fulfills the belief-role specified by the theory, and so is the referent of 'belief', is rational. So too for 'desire'. Hence the conclusion that our attitudes are constituted by their rationality: the basis of Karl's believing and desiring as he does and not otherwise is the fact that it is rational for Karl to do so.[56] It is worth observing the circularity here. Lewis's theory of attitudes is justified by his theory of language, which in turn depends upon his theory of the attitudes. I do not call the circularity vicious: plausibly, we need only grant him a sensible dogmatism about folk-psychology for all his ambitions to be met. But the circularity is there.

Let the spectre of dogmatism lead us onward to our last point. Lewis is concerned almost exclusively with the metaphysics of meaning. 'I am not really asking', he admits in 'Radical Interpretation[57], 'how *we* could determine these facts. Rather: how do *the facts* determine these facts?' But why is Lewis, in contrast to his fellows, so little interested in how we could determine the facts of meaning? I think there are two parts to a good answer here. The first part is simply that, because he says so much about language and the attitudes, he has already done a great deal to show how we could determine the facts of meaning. He tells us, however implicitly, where to look for evidence and how to interpret it.

Yet the last, and one might think the hardest, yard he passes over in silence: what precisely is the nature of our evidence, and is it really good enough to get

[56] Lewis 1994 p 321.
[57] Lewis 1974 p 110.

us the knowledge we're after? I think he chooses to pass over these questions precisely because he is a dogmatist, as he admits in 'Elusive Knowledge'. 'It is a Moorean fact that we know a lot', he says[58], and our knowledge comes in 'all sorts'. The point of that article is to finesse the definition of knowledge in such a way as to acknowledge our fallibility. But Lewis is not inclined to treat the question of how knowledge in a given domain is possible as itself a problem. This is the sense in which I call him a dogmatist, and in this he differs markedly from Quine and to some extent from Davidson too.

Lewis's Principle of Charity may be stated thus: **some sentence S the attitudes of some subject S are just those that A) are substantively rational given S's history of perception and B) rationalize S's behaviour.**

Conclusion

For Quine, the Principle of Charity governs how we should go about interpreting languages; in Lewis' hands, it governs what interpretations of a person's attitudes are correct. The widening of the domain over which charity operates is an intrinsic feature of the story: in Wilson's early formulation, only a fragment of language (proper names) is considered; all those writing after Quine show at least some interest in mental as well as linguistic content. There is less of a clear direction regarding whether charity is oriented towards metaphysics or epistemology: of the authors discussed, Lewis is almost alone in his staunchly metaphysical interest in how 'the facts determine the facts'. The justifications offered for Charity are varied: Quine balances probabilities while Lewis directs us in a neat circle, but I think that Davidson's discussion is most interesting. Charity is plausibly both a general and a specific prerequisite of interpretation. One point of near-consensus, meanwhile, is that Charity is not alethic but

[58] Lewis 1996 p 418.

epistemic. Davidson, who offers the most truth-centric account of all our major figures, is perpetually acknowledging his shortcomings on this score in later reworkings of his ideas.

The movement from proper names to (in principle) the full range of propositional attitudes is likely a natural result of increasing ambition among Charity theorists. Whether interpretation is construed epistemically or metaphysically is, as we have seen, mostly a matter of whether a particular philosopher is a sceptic or a dogmatist. The approaches of Quine and Davidson stand now rather as relics from an age of philosophical heroism, with Charity taken as a lever by which they may turn the world. Lewis is more modest, and appealing to any who have simply been struck by the phenomenon of meaning and seek an explanation for it specifically. It seems clear from past discussion that the best Principle of Charity will be epistemic rather than alethic, with the task being to spell out exactly what epistemic constraints it will invoke. Accordingly, let us turn to two recent attempts at this task: the rationality maximization of Robert Williams, in Chapter 2; and the knowledge maximization of Timothy Williamson, in Chapter 3.

Chapter 2: Williams and Rationality Maximization

Robert Williams advances a theory of interpretation founded on the maximization of rationality.[59] The correct interpretation of an agent, on his view, is that which best rationalizes the agent's dispositions to act in the light of their experience. I will explain the details of his account, as well as the motivations informing it. Then, I will consider potential objections to this account from the perspective of knowledge-first epistemology. Is as internalist a standard as *rationality*, so the worry goes, really adequate to interpretation in all cases? Since Williams is fairly flexible in his approach, allowing that his view can fit different epistemologies, I will find that his account can withstand this critique. Build a bit of externalism into your account of rationality, and the sort of cases which concern externalists can be addressed.

I: The First of Three Tasks

Williams distinguishes three tasks for any theory of interpretation: specify what gets interpreted, what interpretations are, and what makes it the case that a particular interpretation is correct. What gets interpreted, Williams has it, are states of an agent. At a first approximation, these might be thought of as brain states that are classified by type (belief, desire, intention) on the basis of their functional roles.[60] Further, Williams expects these states in turn to exhibit internal structure: beliefs, desires, and other higher-level states will be composed of recurrent elements, which may be thought of as atomic concepts equivalent to words.

[59] Williams 2020.

[60] Williams cites Fodor's language of thought hypothesis as a model here. See Fodor 1975.

Interpretations will map higher-level states to propositional contents, and the elements of those states to sub-propositional contents: individuals, sets, and logical operations, as the contents of words would be specified.

The story about Correctness favoured by Williams is as follows. The correct interpretation of an agent is that which best rationalizes their dispositions to act in the light of their course of experience. Suppose I am disposed to check the BBC Sport website. This would, for instance, be rationalized by a desire to know the football scores and a belief that the website reports those scores reliably: given that pairing of belief and desire, checking the website maximizes my expected utility.

Of course, the desire for anteaters to belch and the belief that checking the website will cause them to do so would equally maximize my expected utility. Nonetheless, my experience of the reliability of the website's reporting justifies the football belief, while my experience of causality tells firmly against the anteater belief. The rationality of desire is a more delicate matter. However one might want to fill in the details (perhaps it is rational to be interested in human excellence, of which sporting excellence is a variety; perhaps it is rational to build solidarity with a community of fans), the football desire seems a good deal more rational than the anteater desire, which might, at best, be minimally rational for a child with an established interest in the animal.

Given how well it rationalizes my disposition to act, a correct interpretation of me would identify some state as my desire to know the football scores and another as my belief that the BBC Sport website is a reliable source for the football scores. At a finer level of detail, it might identify some complex state I entered briefly before checking the website as a *wondering*, one element of which would be identified as my football concept.

II: Structure and Substance

I ought to believe that the BBC Sport website reliably reports football scores; I ought not to believe that checking it causes anteaters to belch. This much rationality demands: so says Williams, who has a high view of rationality's demands. On an alternative view, rationality is merely structural: what matters is that an agent's attitude exhibits the right formal properties, independent of their content. In the first instance, an agent's beliefs must be consistent. Their beliefs, desires, and actions must together exhibit means-end coherence: that is, they must act in ways that they believe will bring their desires closer to fulfillment. At a level of greater sophistication, agents must satisfy the axioms of whatever developed probability and decision theories we might prefer. Within this structure, however, there are no further constraints on the substance of an agent's attitudes.

Williams denies that merely structural rationality is an adequate basis for interpretation for the following reasons. A given agent (Lucy, for instance) has a limited sphere of direct influence and awareness: her 'local bubble'. In Leeds, Lucy cannot hear or see what happens in York, nor can she affect what happens there by her actions.[61] York is outside her local bubble. Not long ago, however, it was within the bubble, as she had been there to view a house she wanted to purchase. At the time, she believed that there was a house in front of her, and desired to own that house. Currently, she believes that there will be a house there next Tuesday, and she desires to sleep in the bedroom of that house Tuesday night. But what about her present-directed beliefs and desires? Here are two interpretations: Standard, on which she believes that there is a certain house in York and desires to own it; and Paranoid, on which she believes that all beyond her local bubble is void, and is indifferent about this.

[61] At least, not without some form of mediation: as Williams notes, the possibility of instant mediation over the internet etc can be eliminated by banishing Lucy several centuries into the past.

Obviously Standard is the correct interpretation. The trouble is that the beliefs and desires attributed by Paranoid seem to be structurally rational. However strange the Paranoid beliefs, they are perfectly consistent. Things are as they appear to Lucy within her local bubble, which changes as she moves, and beyond the bubble lies nothing. Within her bubble, Lucy updates her beliefs on the basis of her changing experience, and satisfies the axioms of probability theory. Lucy also displays means-end coherence. Suppose she signs a legal document. According to Standard, she did so because she believed that this would fulfill her desire to own the house in York. But Paranoid offers a viable alternative explanation: she believes that signing will fulfill her desire to sleep in the bedroom of a house answering some attractive description on Tuesday night. Structural rationality, then, is not sufficient to select the Standard interpretation of Lucy over the Paranoid one. Yet the Standard interpretation is correct. So merely structural rationality is not the basis of interpretation.

What is required for interpretation, then, is substantive rationality. Lucy ought not to believe that the world is void beyond her local bubble, and she ought not to be indifferent whether it is so: not because those attitudes cohere poorly with Lucy's background beliefs and desires, but because they are, in and of themselves, irrational. Ultimately, a fully developed theory of interpretation along these lines will require a fully developed epistemology specifying just what substantive rationality involves. This Williams does not supply: he is offering the outline of a theory of interpretation, to be filled in by first-order epistemology. Williams is happy with this: a fully developed epistemology is something we already want at least as much as a fully developed theory of interpretation, and examining what happens when different accounts of substantive rationality are placed into Williams’s framework affords a new dimension along which epistemologies can be compared.

Note, however, that Williams is making some important assumptions about what the correct epistemology will look like. In order to do the work he wants, an epistemology has to be what he calls intolerantly anti-sceptical: it will imply not only that it is rational to hold certain beliefs about what is beyond one's local bubble (simple anti-scepticism), but that it is irrational not to hold some such beliefs (intolerant anti-scepticism). Consider an interpretation of Lucy (Deviant) on which she is agnostic whether the York house exists now. Deviant also gets Lucy wrong, and if Standard is to be selected over Deviant, then it must be more substantively rational to believe that the house exists than to suspend judgement on the question.

III: Stages and States

Another variation that Williams rejects is a theory according to which what is interpreted are stages. That is, instead of mapping specific states, and even the ur-elements of complex states, to contents, an interpretation might simply map an agent at a time to a set of attitudes. He prefers states over stages for the following reasons.[62] Consider two agents, Smartypants and Blockhead. Smartypants can register finitely many sensation-types, and by complex operations like those which occur within human agents, produce finitely many action-types in response. The relationship between sensation inputs and action outputs can be mapped by a finite function, the Smartypants function, which could in principle be expressed in table-form with every sensation listed on one side and the action to which it leads beside it on the other.

Blockhead too can register finitely many sensation-types and execute finitely many action-types: indeed, his potential sensations and actions are exactly the same as Smartypants. Moreover, exactly the same sensations will produce in Blockhead exactly the same actions as in Smartypants. But the means by which sensation produces action is Blockhead is altogether simpler. Within Blockhead

[62] Williams cites Ned Block (1981) as his source here

is a copy of the table describing the Smarty-pants function. When, say, Blockhead hears someone ask what time it is, sensation-module 1978B activates, which triggers the paired action-module 1023A, and so Blockhead utters ‘It is 10am’.

Smartypants and Blockhead might share the same courses of experience and dispositions to act. Nonetheless, they would not share the same beliefs and desires. Blockhead utters ‘It is 10am’ not because he believes that it is 10am and desires to answer his questioner, but because that’s the outcome his mechanism happens to produce in his present circumstances. He no more believes that it is 10am than a toy cowboy believes that there is a snake in his boot when the string in his back is pulled thus and so. Any theory of interpretation, then, had better predict that the attitudes of Smartypants and Blockhead differ, even given that their outputs and inputs are the same.

Rationality maximization, naturally, would need to find some way in which it is less rational for Blockhead to believe that it is 10am than it is for Smartypants to do so. But given that they have undergone exactly the same courses of experience, this would be puzzling. Perhaps we might favour an epistemology according to which the same total experiences can rationalize different beliefs, but this would be a controversial assumption that is better avoided if possible.[63]

Fortunately, state-based interpretation does make it possible to avoid this assumption. At the very least, a state-based interpretation according to which Blockhead believes that it is 10am and desires to answer his questioner would have to identify two states of Blockhead, one playing a belief-role and the other a desire-role. But Blockhead’s internal processing simply isn’t structured that way. There are no states of Blockhead that play belief-roles or desire-roles, just a

[63] For my own part, I think it worth taking this option seriously. Suppose Smartypants utters ‘I believe that it is 10am’. Given what course of experience would it be rational for Blockhead to do likewise? We would need something like Blockhead being induced to undergo a ‘sensation’ as of introspecting a belief that it is 10am, which I do not consider plausible.

multitude of direct sensation-action pathways. Thus stage-based interpretation elegantly explains the difference between the attitudes of Smartypants and Blockhead without taking on heavy epistemological baggage.

Smartypants, whose internal processing is structured like ours, realizes states that play belief-roles and desire-roles. He has seen the nearby clock, which reports that it is 10am. Thus his course of experience rationalizes the belief that it is 10am. Things have gone better for Smartypants when he communicates with those around him. Thus his course of experience rationalizes the desire to answer his questioner. This belief and desire together rationalize his uttering 'It is 10am'. Supposing we could interpret Smartypants as a model of internally coherent reasoning who happens to believe that it is 9am and desire that his questioner perform tasks an hour late, Smartypants would still believe what he ought not to believe and (probably) desire what he ought not to desire. Thus, on Williams's rationality maximization version of Charity, the correct interpretation of Smartypants will map some state playing a belief-role to the content that it is 10 am, and some state playing a desire-role to the content that his questioner is answered.

IV: Constraining Knowledge

One worry about this view of content might be that it is overly internalist. Rationality, goes the gripe, is too much a matter of what's in the head. But meaning is not just in the head: the facts of reference are more sensitive than that to the way the world is, and the ways that we interact with it, independent of how we think about either. Reference is only properly constrained by an *externalist* standard: something like knowledge, for instance. Such concerns are urged especially by Timothy Williamson, who argues that is indeed knowledge that should be maximized in interpretation.[64] I will consider a range of cases where it

[64] Williamson 2007, p 271.

looks like rationality might not cut it, starting with some proffered by Williamson himself.

Internalism, Williamson complains, makes the relation between reference and epistemology obscure. Some judgement could be rational on many assignments of reference to its component parts; it will be knowledge on far fewer.

Case: *Modest Memory*

> Sarah is reminded of an old acquaintance. 'He had red hair just like that. Always talking about his squash team. Wonder how he's doing now' she thinks to herself. There is exactly one acquaintance of Sarah's, Robbie, who meets this description.
>
> Does Sally refer to Robbie?

Our memories dredge up stray thoughts like this all the time. It will just happen that such sparsely descriptive thoughts arise, absent context or richer identification. Sarah might not be able to pin down when she last saw *his* red hair, or recall in detail an instance of *him* talking about *his* squash team, but nonetheless, the thought comes. Since Sarah's memory is mostly reliable, it is rational for her to assent to its promptings. Once she does, moreover, she will generally gain knowledge. Williamson's point, however, is this. As far as rationality goes, it doesn't matter to whom 'he' refers. Suppose it refers to Bob, presently scrolling through a newsfeed in a café over the road, whom Sarah has never seen. Well, Sarah's memory reports that he (Bob) has red hair, and Sarah's memory is mostly reliable. Sarah wouldn't know that she's thinking about a person she's never seen. She's just following the promptings of her memory, and that's rational.

Knowledge is different. Sally can't know that Bob has red hair, because she has never seen Bob. In fact, Robbie is the only man of whom Sarah is in a position to know that he has red hair and who has both red hair and a penchant for talking squash. So knowledge can constrain the reference of Sarah's use of 'he' to Robbie, and not to Bob. It is less obvious that rationality can do this, since it is rational for Sarah to believe her memory quite generally. So it seems like rationality struggles to get what is evidently the right result: that Sarah's use of 'he' refers to Robbie.

Williamson then increases the pressure with a further case. Some failures of reference, he suggests, can be better explained by the absence of knowledge than the absence of rationality.

Case: *Lucky Brain*

> Take a human brain which has been confined to a vat for the past several decades (though it entered the vat as an adult). It is stimulated in such a way as to undergo an experience as of a tall woman in front of it. As it so happens, there is a tall woman in front of the vat: Sally. It thinks 'She's tall'.
>
> Does 'she' refer to Sally?

In *Modest Memory*, there seemed to be a uniquely suitable referent, Robbie, and the question was whether Robbie was a uniquely suitable object of rational belief. The point of *Lucky Brain* is that there seems to be a uniquely suitable object of rational belief: Sally. There appears to be a woman in front of the subject, and there is one. It's generally rational to believe that things are as they perceptually appear, and so it should be rational for the brain to believe that the woman in front of it - that is, Sally - is tall. On a rationality-maximizing interpretation, the brain's use of 'she' should thus refer to Sally.

However, it doesn't seem like the brain relates to Sally in the right way for reference. It's not receiving any perceptual input from Sally, she has no discernible causal influence over the brain's beliefs, and so on. The brain shouldn't get to refer so easily to beings beyond the vat. So, we may suspect, maximizing rationality gets the wrong result here. Meanwhile, knowledge looks a lot more helpful. The brain's relation to Sally is wrong for knowledge, in much the same way that it is wrong for reference. The lack of perceptual input and causal influence intuitively count against both. The brain cannot express knowledge by 'She's tall' on an interpretation on which 'she' refers to Sally, and so that interpretation would not be correct. A knowledge-based theory of content is superior to a rationality based one here, because it gets the right prediction for the right reasons.

Another way to generate cases of rational belief that intuitively ought not to secure reference is by exploiting the gap between a subject's ability to interpret evidence and what their evidence in fact supports.

Case: *That Many*

> Hardy has no special competence or interest in mathematics. He has made no attempt to learn anything about prime numbers, either by consulting authorities or employing his own computational powers, other than as he was long ago induced to do by his formal schooling. He mutters 'There are that many of them'.
>
> Is an interpretation of Hardy according to which by 'that many' he means 21 and by 'them' he means prime numbers between 100 and 200 'them' thereby more likely to be correct?

The worry is this. It is rational for Hardy to believe that there are 21 primes between 100 and 200, because this proposition is so well supported by his evidence. The basic principles of number theory, which Hardy grasps, wholly justify belief in this proposition. So fundamental is mathematics, indeed, any evidence whatever suffices to justify the entire edifice of number theory. The interpretation according to which Hardy says that there are 21 primes between 100 and 200 credits him as believing what he ought, in fact, to believe.

The fact that Hardy ought to believe does not, in this case, seem to have much bearing on whether he does. We would only expect Hardy to believe this if he can work out that his evidence supports this. But, *ex hypothesi*, he cannot. He lacks the requisite mathematical competence. Without it, the underlying relationship between the proposition supposedly expressed and Hardy's evidence is moot. Again, this would make sense if reference were tied to knowledge. It is only through mathematical competence that Hardy could know that there are 21 primes between 100 and 200. So knowledge is once more a better guide to reference that rationality.

Finally, we might consider putative cases of *unreasonable knowledge*, as defended by Maria Lasonen-Aarnio[65]. The basic idea is as follows. Suppose some agent starts out knowing that P. She then acquires new evidence: a defeater, as it is often described. Apprised of this new evidence, the most rational response is to abandon the belief that P. But the agent ignores this evidence, retaining her belief that P. This belief remains knowledge even though the new evidence renders it unreasonable.

Case: *Red Bag*

> Suzy visits John's party. In his living room are four bean bags. Of the two in the middle, one is red and the other blue. The usually honest and reliable John falsely tells Suzy that there is trick lighting in the room: the one that appears red is beneath a bulb which will make anything look red; the one that appears blue is actually red. Suzy listens politely, but does not give the matter more thought. The next day she reports the scene: 'There were some bean bags. There was *that one*, which was red and in the middle, and I think three more'.
>
> To which bag does Suzy's use of 'that one' refer?

Given the testimony of the typically trustworthy John, Suzy should believe of the actually-blue bag, and not the actually-red bag, that it is red. If the correct interpretation of Suzy is that which maximizes rationality, therefore, 'that one'

[65] Lasonen-Aarnio 2010.

refers to the actually-blue bag. As in *That Many*, however, my judgement is that reference does not 'default' to the presumed most rational belief. The defeater may be part of Suzy's evidence, and for all that is rational to adjust her beliefs accordingly, such adjustment requires concrete cognitive activity. The fact that John has fed her false testimony does not influence her later belief unless she thinks that testimony through, which, *ex hypothesi*, she does not. So rationality maximizing seems to lead us astray once more.

If Lasonen-Aarnio is right, moreover, knowledge does better. Suppose, however, that 'that one' refers to the actually-red bag. Then Suzy expresses a safe belief. That is, Suzy could not easily have been mistaken about whether the actually-red bag was red. Had she trusted John's testimony, she would simply have suspended belief about whether the bag was red: John never denied that it was red, merely undermined the basis of her belief. As Lasonen-Aarnio argues, this belief retains its status as knowledge, given Williamsonian assumptions. If the correct interpretation of Suzy is that which maximizes knowledge, therefore, 'that one' refers to the actually-red bag. Here too knowledge trumps mere rationality.

V: Constraining Rationality

Let's grant that I have identified the correct interpretations in the above cases. The question is then whether the inferior interpretations are really rationality-maximizing ones. In *Modest Memory*, the idea is that it is rational to trust your memory, and so rational for Sally to assent to the passing thought articulated as 'He had red hair', no matter the interpretation of 'he'. That's plausible so far as it goes, but not all interpretations need be *equally* rational. Build even a bit of externalism into your conception of rationality, and the inequality is clear. Yes, Sarah's memory may be reliable in general, but if 'he' refers to Bob the stranger, it has gone badly wrong in this instance. Normally, a memory of someone having red hair would be causally derived from a perception of red hair. If we think that rationality is related to causal pathways, or normal functioning,[66] or that the Bob-interpretaton leaves Sarah with a mere pseudo-memory falling outside of the relevant type of reliable belief-forming process, then the belief that Bob has red hair will be much less rational than the belief that Robbie has red hair.

Much the same applies to *Lucky Brain*. Again, it is somewhat rational for the brain to believe that there is a tall woman before it, as that is how things perceptually appear, but from any strongly externalist perspective, it is not *very rational* to do so. Once again, a usually reliable faculty is misfiring. Still, we might think, all that is needed is for the woman in front of the vat to figure in the most rational interpretation of the brain, so slight rationality is enough. However, strictly speaking, the somewhat rational belief is the belief that there is *a* tall woman in front of the brain. While there is a woman meeting the description 'the tall woman in front of the brain', it is not at all clear that it is especially rational to believe, of Sally specifically, that she is tall. The woman does not causally influence the brain's beliefs about its environment: she simply shares the space without interacting with the vat or its mechanisms at all. Likewise, the brain has no properly functioning faculty, or reliable belief-forming process, trained on Sally. She may be in a special position in regard to the *accuracy* of the brain's belief, at least in the actual case, but rationality is another matter entirely. Williamson vaguely complains that, supposing that the brain can hold beliefs about Sally, then 'there need be no further obstacle to classifying them as justified in the relevant sense'. But there are plenty of plausible obstacles to classifying them so; whichever obstacle we deem decisive, what matters is that the rationality-maximizer need not class them as rational.

[66] In fact, it is Plantinga's account of warrant in terms of properly-functioning belief-forming systems that Williams favours as the basis for an account of rationality. See Williams 2020, Plantinga 1993.

That Many is more delicate. There is a clear sense in which it is rational to believe a mathematical truth: as discussed, everyone's evidence entails every mathematical truth. In practice, though, things are more complicated. A crooked accountant might believe that a sum adds up because it's in his interest that it does, but if he believes on that basis he is irrational even when he believes truly. We should distinguish between *propositional* rationality and *doxastic* rationality. If there are good reasons for believing a proposition, then believing it is propositionally rational. If an agent ignores those good reasons, and believes instead for bad reasons, their belief is doxastically irrational. Grant that doxastic rationality is what we want to maximize, and the issue disappears. Because Hardy is not competent to assess the evidence supporting the relevant proposition, his belief would not be doxastically rational. It could only be a lucky guess. Thus a doxastic rationality maximizing interpretation is adequate in this case.

Appeal to doxastic rationality also handles the final case, though it will take a bit more work specifying just how. The natural reading is that Suzy's use of 'that one' defaults to the actually red bag, in the absence of reflection on John's testimony. Underpinning this verdict, presumably, is the assumption that the basis of Suzy's belief is her veridical perception, and not John's false testimony. It is rational to believe that a chair is red and in the middle on the basis of a veridical perception of a red chair in the middle. On the basis of veridical perception, it is rational for Suzy to believe that the actually red chair is red. It is not rational for her to believe on the same basis that the actually blue chair is red.

Things are a bit more complex than this: Suzy is ignoring a potential defeater, and so her belief that the red chair is red is, as we have seen, less than fully rational. Crucially, however, the irrationality of ignoring John's testimony is baked into the story, a fixed point around which any interpretation of Suzy must turn. The most rational course of action available might have been to change her beliefs on the basis of John’s testimony. But this Suzy did not do. The basis of her belief, as noted, is the perception, and not the testimony. So much cannot be interpreted away. The available interpretations are that she irrationally ignored John’s testimony, and irrationally believed that a blue chair was red on the basis of her perception; or she irrationally ignored John’s testimony, and rationally believed that a red chair was red on the basis of her perception. The latter course is clearly the more rational, even if it is still only imperfectly rational. And so rationality maximization gets the right verdict in *Red Bag*.

While rationality maximization may appear to be vulnerable to externalist concerns, close analysis shows that building a degree of externalism into the account of rationality one uses can secure plausible verdicts in tricky cases, especially when the causal origins of beliefs are taken into account. Rationality as such is not doomed to be insufficiently externalist for constraining reference, contrary to the gloomy prognostications of Williamson. Rationality maximization is an attractive account of content, but its proponents owe some account of how exactly they can avoid the sort of externalist traps that Williamson has laid out for them: at best, I have merely reviewed some of the options available to them. The important point for now is that the options are there. The right way to handle such cases is a topic to which I shall return later.

Chapter 3: Williamson and Knowledge Maximization

Williamson favours an account of Charity that maximizes knowledge. In discussing his view, I will begin by explaining the dialectical context within which it is introduced in *The Philosophy of Philosophy*. Williamson wants to defend the availability of philosophical evidence against what he calls judgement scepticism. To do this, he needs a theory of content on which content is constitutively connected to the environment. Knowledge maximization is such a theory, combining the virtues of two rival theories - alethic and causal - while avoiding their vices. However, I argue that simple knowledge maximization as it stands is flawed, since it takes insufficient account of dispositions.

I: Judgement Scepticism and Philosophical Evidence

Williamson presents knowledge maximization as a magnanimous mean between the pusillanimous acceptance of scepticism and the hubristic assertion of the impossibility of error. The dialectical situation which the proposal addresses is as follows. Williamson wants to allay a certain kind of sceptical worry, what he calls judgement scepticism (and others might call scepticism about intuition).[67] There are mountains, say mountaineers, glaciologists, and Timothy Williamson. Certain revisionary ontologists, such as Ted Sider, say otherwise.[68] Williamson's first line of response is that he *knows* that there are mountains; it's just part of his evidence that, for instance, there are mountains in northern Italy. Likewise, it is part of the glaciologist's evidence when trying to explain glaciation patterns in northern Italy,

[67] Williamson 2007, Chapter 7, 'Evidence in Philosophy'.

[68] Sider 2013. It is worth noting that the 'Sider' of this discussion is primarily the authorial voice of 'Against Parthood'.

and the mountaineer's evidence when planning expeditions there. What's more, it was once part of Ted Sider's evidence before he adopted his snazzy ontological theory, and it was remiss of Sider to ignore this evidence and adopt a theory that it so decisively disconfirmed.

It is at this point that judgement scepticism intrudes. Do we really know, Sider replies, that there are mountains in northern Italy? Is the human mind, in either its philosopher, geologist, or mountaineer varieties, really so attuned to reality's underlying structure? Sider being savvy to Williamson's wiles, he is not likely to appeal to the bare metaphysical possibility of error or the psychological availability of doubt to press this point. He is going to suggest that, according to Williamson's own evidence, the possibility that he might have erred about whether there are mountains is not particularly remote. Not all cultures distinguish features of the earth's surface in quite the way Williamson does. Alien species might make still stranger divisions of the earth's surface. So it is only by chance that Williamson does not say 'the north Italian crust is acutely enfolded, but there are no mountains in north Italy'. Mistakes about mountains are thus very different from the mistakes found in more familiar scenarios of perceptual scepticism: according to Williamson's evidence, the perfect simulations hypothetically delivered to envatted brains are technologically infeasible, and the existence of demons extremely implausible. He is thus obliged to take the no mountains scenario seriously, as he is not obliged to take the BIV hypothesis seriously.[69] As it stands, his supposed escape from relatively easy error is a convenient mystery. Before citing his beliefs about mountains as evidence pertinent to the assessment of ontological theories, Williamson had better dispel this mystery.

It is not merely mountains, of course, which are vulnerable to the assaults of the judgement sceptic. Certain styles of revisionary metaphysical argument call into question our habitual judgements quite generally. What Williamson wants,

[69] Williamson 2007, pp 250-251.

therefore, is an equally general explanation for his many escapes from error. So he tries to establish that there is a general tendency for beliefs as such to be true.[70]

II: Evolution and Content

He begins by appealing to evolution.[71] Surely true beliefs are more conducive to fitness. Say I believe that there is an oasis to my north, and desire to visit that oasis. If my belief is true, I will eventually be able to drink; it is false, I may well die of thirst. Agents who, in general, believe what is true and desire what is good for them tend to prosper. Unfortunately, true beliefs are not uniquely suited to furthering fitness: with the right adjustments, false beliefs can do just as well. Suppose I believe that Question Time is being filmed to my north, and I desire to be in the audience. Then I will end up at the oasis. Say I believe that, by taking a drink, I indicate that I am ready to ask a question. Then I will stave off dehydration. There are many different possible combinations of beliefs and desires, where an agent believes what is true* and desires what is good* for them, that will enable the agent to prosper.

To see this, we make an arbitrary mapping of propositions, and syntactical elements of propositions, taking any proposition P to ^P. Then we define 'true*' and good*' as satisfying these equivalences: that P is true* if and only that ^P is true, and that P is good* for an agent if and only ^P is good for an agent. Suppose that an agent desires that P, and believes that action A will bring it about that P: in this case, P is that I am in the Question Time audience, ^P is that I am at the oasis, and the action is walking north. So the agent performs action A, that is, I walk north. Now suppose that my beliefs are true*. By our definition of 'true*', the proposition that ^(if I do A, then P) is true, since we have granted that if I do A, then P is true*. Which is to say, because it is true* that, if I walk north, then I will

[70] Williamson 2007, Chapter 8 part 1.
[71] Ibid, Chapter 8 part 2.

be in the Question Time audience, it is true that, if I walk north, I will be at the oasis. Now further suppose that P is good* for the agent: by the definition above, ^P is good for them. Desiring that P (that I am in the Question Time audience), which is good* for me, I walk north in the true* belief that this will bring it about that P. This action brings it about that P^ , which is actually good for me. In this case, what is good for me is being at the oasis, and it is brought about by my odd Question Time related beliefs and desires. But it is in the nature of the permutation that whenever I believe what is true* and desire what is good*, I end up bringing about what is good for me, at least subject to the ordinary exigencies that plague trying to bring about what is good for me because I desire what is good for me and have true beliefs about how to achieve it.

Bringing us back to evolution, doing what is good for one and one's offspring will contribute to fitness. By going to the oasis, our remote ancestors were able to survive and reproduce. But as we have seen, truly believing that walking north will bring them to the oasis, and desiring the genuine good of being at the oasis is not crucial for fitness, since true* beliefs and good* desires achieve the same result.

This point is properly speaking the beginning of Williamson's argument for knowledge maximization. Granted that our ancestors could have got by with true* and good* desires, it remains the case that they did not. That they did not, moreover, is no mere quirk of fate: there is some robust sense in which they could not have done so.[72] On the interpretation on which they do have true* beliefs and good* desires (Int*), our ancestors' propositional attitudes are radically disconnected from their actions and environment. That's wrong. Any agent's propositional attitudes are connected with their actions and environment, and this connection is constitutive. The important question is the nature of this constitutive connection.

[72] Though perhaps, as Williamson concedes, falling short of full metaphysical impossibility.

Williamson considers two main approaches here, causal and alethic. One way of spelling out the absurdity of Int* is as follows: Question Time is a political panel show in our ancestors' far future. It does not impinge on them in any way. How could it ever have entered their heads? The concern here is that an agent can only holds beliefs about an object if that object is causally connected to them; we think about objects because they force their way into our heads. The constitutive connection between propositional attitudes and the environment, on this view, is causal.

However plausible such a proposal might seem initially, however, Williamson finds it unsatisfactory. First, not all beliefs seem to require causal connections: mathematical beliefs appear to be about abstract objects causally unconnected with us. Second, and more crucially, not all causal connections seem relevant to reference. Suppose my computer login screen shows a picture of an Alpine valley. Seeing it, beliefs about the valley come naturally to me. But beliefs about the glaciers that shaped the valley do not, and nor do beliefs about the liquid crystal by means of which the valley is displayed. I am in the right kind of causal relationship with the valley, but not the glaciers or the crystals. But what makes for an appropriate causal relationship here is obscure.

Here is another way to spell out what's wrong with Int*. Question Time was not being filmed north of our ancestors. The technology did not exist. Nor was there a British Broadcasting Corporation to make the programme, nor a British parliamentary system for the programme to discuss. If our ancestors were getting Question Time that wrong, was it ever really Question Time that they were thinking about at all? This time, the worry is that an agent can only hold false beliefs about an object if they already possess a background of true beliefs about that object. The constitutive connection between propositional attitudes and the environment, on this view, is alethic: our beliefs tend to be true.

Again, Williamson is sympathetic without being convinced. A bare principle of truth maximization would rule out the possibility of massive error on principle, which Williamson finds implausible. A recently envatted brain, he reasons, would believe great tissues of falsehood. Davidson was moved fancifully to speculate that no one had ever believed that the earth was flat, since one could not go so wrong while still talking about this very earth. Truth alone, then, does not seem to be the right connection between thought and world either.

III: Knowledge Maximization

Williamson's own proposal is introduced using a case along these lines.[73] Suppose I decide that I can infer life stories from the faces of strangers. I look upon one woman, Elsie, and think: She's dark-haired, a nurse, of mixed Greek and British ancestry', and so on. As luck should have it, just behind her and beyond my sight is another woman, Imogen, who exactly matches my description.If the constitutive connection between belief and the environment is alethic, then we should expect my SHE-thought to refer to Imogen, since on that interpretation, my belief is true. Surely, however, this is wrong. I could not see Imogen, and was looking at Elsie throughout: I must have been talking about Elsie.

At this point, it would be tempting to invoke causation again: I must have been thinking about Elsi*e because I was causally connected* to her. But the case can be finessed so that the problems of specifying the right causal chains recurs. Suppose Imogen had been a nurse when Elsie received plastic surgery: she is thus causally connected with my beliefs, insofar as she helped to sculpt the features that trigger my strange guesses. Nonetheless, I cannot have been thinking about Imogen; the causal connections are of the wrong kind.

[73] Ibid, pp 262-264.

So why is it I refer to Elsie rather than Imogen? Because I was in position *to know* about Elsie and not Imogen. I could see Elsie clearly, and this enabled me to acquire knowledge about her: that, for instance, she was dark-haired. But I have no means of knowing about Imogen, however many thoughts I have that would be true were she their subject. So it is to Elsie rather than Imogen that my SHE-thought refers: on this interpretation, I have fewer true beliefs, but more knowledge.

This view respects the motivation behind the alethic proposal: an agent's propositional attitudes are indeed connected with the agent getting their environment right. But an agent has to get their environment right *in the right way*: what matters is knowledge of their environment. Absent the right observations and deductions, no one would be in a position to know that the earth is round, so prior to those observations and deductions, they did in fact believe that it was flat. Knowledge maximization thus improves upon mere truth maximization.

It also respects the motivation behind the alethic proposal: often, we know about objects because we stand in causal relationships to them. But not every causal connection yields knowledge, just as not every causal connection yields reference; sometimes we can have knowledge without causal connections, just as we can have reference without causal connections. Thus knowledge maximization also improves upon causal theories of reference.

Notoriously, Williamson closes *The Philosophy of Philosophy* with an epilogue criticising current standards of philosophical practice. Among his complaints is that crucial claims are vaguely stated. It is thus rather ironic that the knowledge maximization thesis, which is the crucial claim of the book's last chapter, is not given much by way of precise elaboration. Following the example set by Williams's account of representation,[74] we should offer an account of Framing for knowledge maximization, and also an account of Correctness. Giving the Framing

[74] Williams 2020, p 12-13.

means saying what sort of thing interpretations are: what gets interpreted, and what does an interpretation ascribe? Addressing Correctness involves saying what makes it the case that one of the various interpretations available is the right interpretation for a particular agent.

I take it the best Framing for knowledge maximization follows William's Rationality Maximization in mapping specific states to contents. Belief states (on Williamson's view, knowledge-like states) will certainly be mapped to propositional contents; ideally, the constituents of such states would also be mapped to finer-grained contents. This enables greater theoretical flexibility, as will be discussed later. The story about Correctness is that the right interpretation of an agent is that which attributes them with the most knowledge-states. Suppositionally, I was in a belief state SD, associated with the thought 'She's dark-haired'. There are two rival interpretations of me, differing only in the contents with which they map the belief-states I formed after seeing Elsie associated with a series of thoughts using 'she', *E* and *I*. On *I*, SN is a mere belief; on *E*, it is a state of knowledge. There are no other belief states of mine that *I* counts as knowledge but *E* does not. According to knowledge maximization then, *E* is the correct (or at least, superior) interpretation.

Knowledge maximization: the correct interpretation of an agent is that which attributes that agent with the most knowledge-states.

IV: Ties and Channels

Martin argues that there are some cases where two interpretations can be tied so far as knowledge maximization goes, but in which reference is nonetheless clear.[75]

[75] Martin 2009.

The knowledge maximiser should explain how reference is settled in such cases despite the seeming tie. Suppose Lucy is looking at two exactly similar dolly mixtures, one to the right and the other to the left. She thinks ‘That’s pink’, and no more. She can see perfectly well, of each dolly mixture, that it is pink. Whichever ‘that’ refers to, therefore, ‘That’s pink’ articulates knowledge. Since she has no other thoughts, ‘that’s pink’ is the only item of knowledge expressed. So Lucy knows exactly as much whether her use of ‘that’ refers to the left or the right mixture. There is a tie, so far as knowledge maximization is concerned. Yet surely reference need not be indeterminate in such a case. It is perfectly possible for Lucy to look at two different dolly mixtures, but single out one in thought, thinking of the left mixture, and only of the left mixture, that it is pink. How can this be, within the bounds of knowledge maximization?

Williamson’s answer is as follows.[76] There is some definite mental process, occurring at a non-intentional level, by which Lucy singles out the left mixture even while receiving perfectly good visual information about both mixtures. So much is part of the story; something is going on to settle the reference, what matters is whether knowledge maximization is sensitive to it. This process opens up a mental file, in which beliefs are generated by visual stimulus deriving from lefty and expressed using ‘that’. In fact, the only belief generated fails to discriminate between the two, in the sense that it could be knowledge of either, but that belief is still causally related to the left rather than the right mixture. This connection is such that if Lucy were asked is that on the right, she would say yes. In that case, of course, knowledge maximization would have a clear winner. In the actual case, we might say that Lucy’s use of ‘that’ is a *channel for knowledge* of the left and not the right mixture.

Granted, knowledge maximization as stated makes no reference to counterfactual knowledge, nor channels for knowledge. This gap is bridged by an appeal to *naturalness*. Reference is a natural relation. In worlds close to the case world,

[76] Williamson 2009, Reply to Martin 2.

Lucy’s use of ‘that’ encodes most knowledge if it refers to lefty. So at those worlds, it refers to lefty. Now, suppose it refers to righty at the case world. Then, there is a strange gap between reference at the case world and at close worlds. If her use of ‘that’ refers to lefty at the case world, this gap disappears. Such gaps are unnatural; reference is a more natural relation if the correct interpretation of Lucy's use of 'that' is consistent between close worlds and the case world. Thus the correct interpretation of Lucy is that she is referring to the mixture on the left.

The apparent tie is broken in a way that is ‘consistent with the principle of knowledge maximization’.[77] ‘Even if no knowledge actually happens to be gained’ through a particular channel for knowledge, the very fact that the channel has opened up means that there is now more knowledge that *would* be gained, were the channel used. This settles what the correct interpretation would be in those cases, and ‘the naturalness of the reference relation may still keep the reference constant between the actual case and counterfactual cases in which knowledge of lefty is gained through the channel’.

So, a fuller statement of the knowledge maximization view might be as follows.

1. The correct interpretation of an agent is that which attributes that agent with the most knowledge-states.

2. If there is no uniquely knowledge maximizing interpretation, then the correct interpretation of an agent is that among the actually knowledge-maximizing interpretations which attributes the agent with the most knowledge-states at close possible worlds.

[77] Ibid.

V: Williamson and the Three Tasks

Such, in outline, is the knowledge maximization theory of content. Per Williams, there are three tasks for a theory of content to fulfill: specify what is the direct object of interpretation, what interpretations are, and what makes it the case that a particular interpretation is correct (2.1). Williamson concentrates on the Correctness task; while I have suggested an answer for the Framing task (3.III), it is now time to consider the remaining two tasks in greater detail.

What exactly is the direct object of interpretation on the knowledge maximization theory? Williams, as we have seen, opts for the states of an agent, mapping those states and sub-states to contents of appropriate granularity: propositions for belief-states and so on down the line. But we might also choose the agent at a time, mapping the agent at that time directly to a set of attitudes without regard for the agent's states: or maybe the agent over the course of their life, mapping the agent to a larger set of time-indexed attitudes; or some other possibility. The only point on which Williamson is really explicit is that individual concepts are not the object of their own singular interpretations, with a full picture of what a person thinks being built up only gradually as we interpret each of their concepts successively. Rather, interpretation is properly concerned with 'the subject's total system of thoughts'.[78] Although he doesn't directly contrast the agent with their system of thoughts, I believe that the system of thought is a better candidate than any alternative. To see why, we should examine a problem Williamson examines in passing but for which he does not give an entirely satisfactory solution.

[78] Williamson 2007 p 259.

If interpretation maximizes knowledge, how is massive error possible? Williamson answers that error usually entails ignorance, and massive ignorance is to be expected. Since the relevant principle is to maximize knowledge, rather than minimize ignorance, there is no problem. This answer, I think, is not especially helpful. We know that Williamson knows that there are different ways in which one might fail to know. For any true proposition P, a subject might 1) believe P and know that P; 2) believe P and, as in a Gettier case, fail to know that P; 3) believe not-P; or 4) believe neither P nor not-P. Williamson's general concession to the scale of human ignorance does nothing to explain how applying the principle of knowledge maximization could generate cases of the 2nd and 3rd kind, where we positively ascribe ignorance. Since these beliefs fail to constitute knowledge, an interpretation which attributes such beliefs to a subject does not maximize that subject's knowledge compared to an interpretation which does not. And yet all human subjects hold ignorant beliefs. Williamson himself claims to have generated Gettier cases with his own lectures.[79] Why should he believe this, when an interpretation on which his audience does not hold a Gettier belief would surely be equally knowledge maximizing?

The answer, I take it, is that interpretation is constrained by the overall structure of the agent's system of thoughts. To see how this might work, let us look at the linguistic case. Take these sentences of a language with the logical vocabulary of English, but novel predicates:

1) All bloofs are verns.
2) All verns are miggles.
3) Some bloofs are not miggles.

Suppose Tom has spoken these three sentences, and we are trying to interpret these utterances. We aim to maximize his knowledgeable assertions, but soon get stuck. For the three sentences he has asserted are jointly inconsistent. If all bloofs are verns,

[79] Ibid, 192.

per 1); and verns, miggles, per 2); then all bloofs are miggles, contra 3). So we are forced to interpret Tom as asserting ignorantly at least once. Or more specifically, we are so forced if we mean to interpret ‘bloofs’ ‘verns’ and ‘miggles’ consistently across all three utterances: otherwise we might pair the utterances with any propositions Tom happened to know. But in the process of linguistic interpretation, we assume that the meanings of sentences depend systematically on their syntactic structure and the fixed meanings of recurring subsentential parts: language is compositional. A full interpretation of Tom’s utterances would interpret the subsentential parts of his sentences, and crucially the novel predicates ‘bloof’, ‘vern’, and ‘miggle’. And it is in assigning fixed interpretations of the subsentential parts of Tom’s speech that we end up constrained to attribute ignorant utterances to him.

This is exactly the sort of constraint knowledge maximization requires. Ordinary humans make ignorant assertions, and Tom is an ordinary human. The question is how similar constraints might generate ignorant beliefs. The answer is that thought, like language, is compositional. Beliefs, like sentences, have recurring components: concepts. The content of beliefs systematically depends on the concepts which compose them. Assigning fixed interpretations to a person’s concepts thus constrains the interpretation of their belief-states. In a simple case, Tom holds inconsistent beliefs, which he expressed through sentences 1-3 above, with concepts corresponding to the logical vocabulary and the three predicates ‘bloof’, ‘vern’, and ‘miggle’. There is no interpretation of all the concepts employed such that all the beliefs composed by them are knowledge. As a fallible human being, Tom’s cognitive architecture is bound to exhibit structures like this: no consistent interpretation of his concepts will count him as believing only when he knows. Knowledge maximization succeeds in generating ignorant beliefs when the direct object of interpretation is an agent’s whole cognitive system, including many belief-states and the recurring concepts which compose them.

That in turns yields an answer to the second outstanding question for knowledge maximization: what is an interpretation? As for Williams, an interpretation is a mapping from the belief-states of an agent to propositional contents, and from the recurring components of belief-states to subpropositional contents: individuals, sets, logical operations, and so on.

VI: Unsafe Reference

Before pressing my own objections to this view, I would like to review a hard case that Aidan McGlynn has offered.[80] It is constructed along somewhat similar lines to my *Red Bag* case (2.IV), in that it involves a veridical perception that is epistemically sub-par in some respect. Reference is supposed to follow veridical perception notwithstanding the way in which it is sub-par. *Red Bag* could be seen as supporting knowledge maximization, since the knowledge maximizer could claim that the veridical perception yielded an instance of knowledge that failed to meet some further condition. McGlynn, on the other hand, argues directly that the veridical perception in his scenario fails to generate knowledge on the basis that it is unsafe. Given that the unsafe perception yields reference, this is a problem for knowledge maximization.

Case: Hallucinogen

> Lucy has taken a potent hallucinogenic drug. In front of her, she seems to see several people. In fact, there is only one person in front of her: Helen. None of the experiences as of people before her seem any more or less genuine to Lucy than the others. Nevertheless, Lucy happens to focus her attention in Helen's direction,

[80] McGlynn 2012a.

> attending to experiences caused by light which has reflected from Helen. Lucy thinks to herself 'She is tall, she is freckled, she has green eyes' etc. Helen is tall, freckled, has green eyes, etc.[81]

According to McGlynn, Lucy succeeds in referring to Helen with her use of 'she'. He also contends that, so interpreted, the beliefs Lucy expresses by her use of 'she' fail to constitute knowledge. This is because Lucy could easily have focused her attention slightly differently, and formed relevantly similar beliefs on the basis of experiences caused by the drug: this might have led her to mutter, for instance, 'she is short', when there is no person who is short in front of her. Thus Lucy's beliefs violate the following plausible constraint on knowledge: S only knows that P if in all nearby possible worlds in which S forms a relevantly similar belief on a relevantly similar basis as S forms her belief that P in the actual world, that belief is true. That is, Lucy's beliefs are not safe, and so, not knowledge. This implies that Lucy's reference to Helen is not explained by the fact that reference to Helen maximizes Lucy's knowledge. So knowledge maximization is not an adequate theory of reference.

This way of putting the matter, I think, overstates the problem for Williamson. Note that McGlynn only argues that the correct interpretation in *Hallucinogen* is not a knowledge maximizing one; he does not argue that there is some knowledge maximizing interpretation in *Hallucinogen* that is not correct. So in fact we can assimilate this case to the Pinky case: straightforward knowledge maximization does not deter a unique correct interpretation, since there is no one uniquely knowledge maximizing interpretation of Lucy's use of 'she'. As we have seen, Williamson accepts the possibility of such ties, and proposes a solution for resolving them: if no interpretation uniquely maximizes actual

[81] McGlynn adds that Lucy knows that she has taken a hallucinogen, and thus possesses a defeater for her beliefs about Helen. However he acknowledges that, given Williamson's views, the issue of safety is more pertinent, and since we have already discussed defeaters in *Red Bag* (2.IV), we shall concentrate on safety here.

knowledge, then we turn to the interpretation that maximizes knowledge at close possible worlds. In this case, Lucy would know ‘she is freckled’ etc were she not hallucinating, given that ‘she’ refers to Helen. This interpretation of Lucy’s use of ‘she’ which maximizes her knowledge at close worlds, and so the correct interpretation at those worlds. Hence it is actually the correct interpretation, because reference is a natural relation which does arbitrarily shift between close worlds.

McGlynn is aware of the manoeuvre, and replies that, while he is unable to provide a counter-example for it, the view as adapted to meet such cases is no longer one on which knowledge comes first in the theory of reference.[82] This may sound surprising: the theory as stated is indeed one on which reference is entirely explained by knowledge, either actual or closely possible. But McGlynn’s point is about explanatory power. We may be able to give an account of *Hallucinogen* in which it is consistent with the theory of knowledge maximization, but, given the weakening of the theory involved in handling ties, it is not clear whether this is a compelling explanation of referential success in this case. In particular, we might wonder whether we have made any advance on the bare claim that reference Lucy refers to Helen because she has the right sort of causal connection to her. To make this more pointed, consider that on any view of content, referring to Helen is going to be a precondition on knowing about Helen: without reference to her, it would not be *Helen* that was the object of Lucy’s knowledge. So it might seem uninteresting that Lucy’s use of ‘she’ to refer to Helen opens up a channel for reference about Helen: *any* instance of successful reference opens up a channel for potential knowledge about the referent. So Williamson’s invocation of possible knowledge does not explain Lucy’s reference to Helen: this is explained otherwise, plausibly through a causal connection, and it is in fact this already successful reference that explains Lucy’s closely possible knowledge of Helen.

[82] Ibid.

This objection, however, is easily resisted. Consider again Imogen, who in 3.III was causally responsible, in virtue of her work as a nurse, for the facial features which prompt the beliefs that I form while looking at Elsie. This, as we say, is the wrong kind of causal relationship: forming beliefs while looking at Elsie, I do not refer to Imogen. What matters is whether knowledge maximization can explain this difference between *Hallucinogen* and the Imogen/Elsie scenario. If it can't, then McGlynn is right that knowledge maximization is in trouble. But it can explain the difference. Perhaps it is, in a very broad sense, possible for me to gain knowledge of Imogen, using 'she' in the same way, in this scenario. She is on the same street, after all, and I could easily go and see her too. Though it would be a very unnatural way to think, I could in doing so continue forming beliefs under the same 'she' mental file as I opened while looking at Elsie, making no distinction between the 'she' I had thought of a few minutes earlier and 'she' I am thinking of now. But this altogether a remote possibility: in close cases where I gain knowledge of Imogen, I open a new mental file for her, and gain no knowledge under the old file.

Contrast *Hallicinogen*. Just as I could go on from Elsie to look at Imogen, so Lucy's drug could wear off and she could go on to see Helen without its influence. In this case, she would not open a new mental file. She would continue adding beliefs, using 'she' in the same way, to the stock acquired while still hallucinating. These beliefs would be knowledge. Thus the possibility that Lucy gains knowledge of Helen using 'she' is indeed closer, and far closer than the possibility of my similarly gaining knowledge of Imogen in the other case. So invoking knowledge does appear to be genuinely explanatory. The difference between the causal connection in the Imogen/Elsie case, and the causal connection in *Hallucinogen*, is that the latter makes for much closer possible knowledge, and this is far more informative than saying merely that the connection is of the right kind.

It is also worth noting that we are seeking a general theory of content, and we should assume that every instance of successful reference depends on as obvious a causal connection is those we have considered here. Suppose I am explaining the line of succession in the United Kingdom: ‘the next child born into the succession - Princeling, let’s say - will enter it ahead of anyone in the previous generation who is already behind Princeling’s royal parent’. The unborn Princeling is evidently not standing in any causal connection to my beliefs about them, and yet I seem to make successful reference. Perhaps a causal theory can be nuanced to explain this - I am primarily aiming to defend an interpretationist theory, rather than to attack a causal one - but a knowledge theory straightforwardly predicts this. There is much that I can know about Princeling, qua arbitrary member of UK line of succession, and indeed qua arbitrary human born in 2023. More than this, there is some uniquely identifying knowledge I possess about Princeling: that they will be the next child born into the succession. So the knowledge theory explains my reference to Princeling, Jane’s reference to Helen, and it explains my failure to refer to Imogen. It is obviously an advance on invoking the right sort of causal connection.

Still, I think that there is more pressure for us to put on Williamson here. In particular, there is at least one significant disanalogy between *Hallucinogen*, in which Lucy sees only a single real person, and Martin’s earlier case, in which a Lucy sees two dolly mixtures. In Martin’s case, there are two ways that we could interpret Lucy as expressing knowledge with ‘That’s pink’. The issue was that there was a small group of candidate interpretations according to which Lucy had the same positive amount of knowledge. Close possible knowledge was invoked to break a tie between a few equally good interpretations. The point of *Hallucinogen*, on the other hand, is that there is no way to interpret Lucy as expressing knowledge. All possible interpretations are thus equally bad at maximizing Lucy’s actual knowledge. Is it really adequate to shrug our shoulders, say that a tie is a tie, and carry over the close possible knowledge strategy from *Pinky* without further elaboration?

I believe not. After all, one of the uses to which Williamson wishes to put his theory is in explaining certain expected reference failures:[83] recall that this was the point of *Lucky Brain* (2.IV). An envatted brain cannot seem to see a woman front of it, say ‘She’s tall’, and succeed in referring to a woman that just happens to be in front of its vat at the time. This is because the brain is ignorant of the woman, or so Williamson says. But if reference is *sometimes* possible in cases of actual ignorance, then Williamson owes us an account of exactly when merely possible knowledge suffices for reference, and when it does not. Presumably the idea of closeness will be crucial here. But he had still better have some satisfactory answer to the question of why the case where Lucy hasn’t taken the Hallucinogen and sees Helen is relevantly close to *Hallucinogen*, but the case where the brain is embodied and sees the beanbag in front of it is not relevantly close to *Lucky Brain*.

The answer is not likely to be ‘Re-embodying a brain is hard’, for consider an amendment of *Hallucinogen*, *Hyper-Hallucinogen*, in which the drug Lucy takes is not merely a hallucinogen, but also has the property of screening-off the vision centres of the brain from optical inputs entirely. I take it that in this case, where Lucy’s person-seemings are entirely caused by the drug, and not at all by Helen, reference should fail just as in *Lucky Brain.* Why should tailoring the drug in this way destroy reference? As with the hard cases earlier presented for Williams’s rationality maximizing view, this is not supposed to be a decisive objection to Williamson or knowledge maximization as a project. I am merely illustrating an outstanding problem for views of this sort to address.

VII: Stability and Disposition

Now I am ready to introduce my own concerns with knowledge maximization as it stands. In my view, the focus on actual knowledge allows an agent’s content to

[83] Williamson 2008, p 271.

slide a little loose from their dispositions. The high-level intuition here is that content should vary with dispositions, and so long as dispositions remain stable, content should be at least somewhat robust to changes in the world. I explore this idea below with examples of two kinds of disposition: recognising natural kinds, and distinguishing the real from the fake. In scenarios where these dispositions are sufficiently masked, the knowledge maximizing interpretation will ignore them. But if the dispositions are present behind the mask, then I submit that interpretation ought to account for them.

Case: Swans

> During Beatrice's life, Europeans discover Australia and document its fauna. Early in her life, she assents to 'All swans are white'. Having heard reports from Australia, however, she smoothly switches to dissenting from 'All swans are white'.

Initially, Beatrice has a belief state which she expresses by the sentence 'All swans are white'. If this belief state is knowledge, then she will be more knowledgeable overall. So the interpretation of Beatrice's cognitive system which maximizes her knowledge at this point is one on which that belief-state is true. In turn, this means that the knowledge-maximising interpretation of what we might call her swan-concept is one on which it does not refer to black Australian swans. Given the other ways in which she uses the concept, for instance, in response to seeing European swans on a river, there is an obvious candidate for the reference of the concept: European swans. Plausibly, the inductive evidence is indeed enough for her to know of European swans that all are white, and of course she will encode knowledge with it during particular riverbank sightings.

Later in life, she has a different belief, which she might express as ‘Not all swans are white’. Since (let us assume for simplicity) all European swans are white, this belief is false if her swan-concept refers exclusively to European swans. On a knowledge maximizing interpretation of the older Beatrice’s cognitive system, this concept will have another reference. Again, there is a clear candidate: all swans, as standardly classified in modern biology, both European and Australian.

A straightforward application of knowledge maximization to Beatrice’s cognitive system yields the result that she has different swan-concepts, or different contents for her swan-concepts, over time. But this, I submit, is the wrong result. Given a sufficiently smooth and swift transition to acceptance of ‘There are black swans’ and other such sentences on the part both of Beatrice and the wider linguistic community, without any particularly deep reflection about what a swan really is, or what is the purpose of their classifying things as swans, then the natural way to describe what happens is that Beatrice used a recurring swan-concept with a stable content to express different beliefs in the light of new evidence. She and the travellers on whom she relies might even say outright that their past belief that all swans are white was mistaken. Of course, we shouldn’t assume that Beatrice or any other agent has perfect introspective awareness of their mental content,[84] but on the face of it that does just seem like an accurate description of the situation, besides which such a comment would express yet another belief that the knowledge-maximizer must factor into their interpretation.

One potential solution is to maximize not knowledge at a given time, but knowledge over an agent’s lifetime. On the European swan interpretation, not only will Beatrice’s ‘Not all swans are black’ belief be false, but all the other swan-beliefs she forms in response to Australian reports will be so too. She will hold, for instance, a false belief expressed as ‘There are swans in Australia’. All

[84] Williamson in particular is likely to deny this in line with the argument of his 2000 *Knowledge and Its Limits.*

these beliefs are knowledge on the generic swan interpretation, so the loss of the initial ‘All swans are white’ belief is more than compensated for. Therefore knowledge maximization can generate the correct result that Beatrice had always referred to swans in general.

The first problem with this response is that it restores the agent as the direct object of interpretation. What gets interpreted is an agent over the course of their life. But as I have already explained there are technical reasons why interpretation should not be framed around agents: only the detailed structure of the cognitive system offers the resources to constrain interpretation properly. The cognitive system as a whole, meanwhile, could change (for instance, through the addition of new concepts) while specific concepts within it remained stable. So in general I doubt that maximizing knowledge over time is likely to be the solution.

Furthermore, there is a vivid way to argue this point directly. Suppose Beatrice never had the chance to learn about Australian swans. She died of a sudden illness before the first reports came back. Tragic Beatrice, as we might call her, uses her swan-concept in just the same way as the younger ordinary Beatrice. The same reasoning that initially swayed us to interpret the younger Beatrice as referring only to European swans also applies to tragic Beatrice. ‘All swans are white’ will count as knowledge for her on this interpretation, but not if she is referring to all swans, and so reference to European swans will maximize her knowledge. Since she never adopts the beliefs that will change this calculus for the ordinary Beatrice, this interpretation will still maximize her knowledge over time.

My concern is that this leaves the younger ordinary Beatrice and the tragic Beatrice with differing contents for their swan-concepts. But the younger Beatrice and the tragic Beatrice are, at this point in time, indistinguishable. They use their swan-concept in the same way in response to the same stimuli. Given how much between the two is consistent, surely the content of their swan-concepts should also be consistent. The fact that ordinary Beatrice lives to hear about Australia doesn't seem like the sort of thing that should cause such a large difference in content between the two.

One point that I want to draw attention to is that both the younger ordinary Beatrice and the tragic Beatrice both share the dispositions that the former goes on to manifest. It is built into the way that the tragic Beatrice uses her swan-concept that, should she hear reports about black swans from the far corners of the world, she would revise her belief that all swans are white, just as the ordinary Beatrice will do. These dispositions are consistent between the two Beatrices, and across the ordinary Beatrice's longer life. Where the consistent dispositions lead, I suggest, consistent content follows. Because the relevant disposition fails to manifest before a certain time, knowledge maximization cannot take it into account. But, I take it, the moral of Beatrice's story is that interpretation should take such masked dispositions into account. Early in her life, Beatrice would recognise an Australian swan under her swan-concept, and so that concept always referred to Australian swans. Insofar as knowledge maximization fails to predict this, it is flawed.

Case: Barns

> Rogier is travelling through a foreign country for the first time. This country has distinctive barn architecture. It also has many fake barns, which employ the same style of architecture. Rogier is so struck by the distinctive architecture he coins a general term, 'barnia'. On the first, day he sees only fake barns, and applies his term when he sees them. The next two days, he sees only real barns, and also applies the term when he sees them. The

> fourth day, he starts to examine some of the distinctive structures that he has been seeing, and finds that some are genuine barns and others are mere barn facades. On examining his first facade, he immediately thinks 'This is not a barnia', and from that point on takes care to determine whether a given structure is a genuine barn or a facade when he uses the term . Eventually, his journey loops back around to its beginning, where he says of the structures he saw on the first day 'none of these were barnias after all'.

There are three potential interpretations of 'barnia' we will want to consider. On the first, it refers exclusively to barn facades designed in the distinctive style of the country. On the second, it refers to real barns so designed. On the third, it refers indifferently to barns and to facades that share the distinctive design. Which is the correct interpretation at the end of the first day? Well, presumably Rogier believes something like 'All barnias are barns'. This will be knowledge on the second, real-barn-only interpretation, and not on any of its rivals. The trouble is that Rogier has many other beliefs that will be knowledge on those other interpretations, but not on the second. Rogier believes, for instance, that he saw 16 barnias that day. For each of those sixteen supposed barnias, he has detailed perceptual beliefs, as well as beliefs locating them in space, as well as locating his encounter with them in time. Not all of Rogier's beliefs can be right: either some barnias are not barns, or Rogier didn't see any that day. Surely the mass of beliefs about specific barnia-sightings outweighs the single general belief about the nature of barnias. Rogier knows more if he really has seen 16 barnias, than if he is right about his universal generalization. So knowledge maximization suggests that Rogier's use of 'barnia' does not refer exclusively to real barns at the end of the first day.

Presumably the indifference interpretation is the best: if the term refers exclusively to facades, then the presence of real barns in a similar style which Rogier can't yet distinguish would seem to threaten his knowledge.[85] Any doubt about this is dispelled by the end of the third day. 'I have seen many beautiful barnias over the last three days', for instance, will only be true if both genuine barns and facades fall into the extension of the term. Our problems start on the fourth day, when Rogier begins to discriminate between the barns from the facades, and, moreover, registers the distinction between them in his use of 'barnia'. Rogier now believes, of at least some facades in the relevant style, that these are not barnias, and this will only be knowledge if the term does not refer to those facades. Moreover, Rogier will presumably revise his prior beliefs in the face of this new evidence: in this case, setting aside the detailed beliefs about barnia sightings we wanted to squeeze knowledge out of earlier. By the end of the trip, the winner is clear. The broad pattern of his barnia-beliefs is that of believing that something is a barnia if he knows it to be a real barn in the relevant style, believing that it is not a barnia if he knows it to be a facade, and suspending judgement in other cases. And, of course, he retains the generalization he believed at the start, that all barnias are barns. At this point, the real barn only interpretation decisively maximizes Rogier's knowledge.

The question is what happens to Rogier's content between the first and last day. Did the meaning of 'barnia' change as Rogier learned more, preserving his knowledge from day to day, or did the term have a consistent meaning that left Rogier badly mistaken in the beginning? I believe the latter. It is part of the story that Rogier engaged in no higher-order reflection about the appropriate use of 'barnia'. On seeing the facade for what it was, he immediately recognised it as belonging outside the category of barnias. This division of conceptual space was part of how Rogier used the term from the beginning, as evidenced by his general belief that all barnias are barns. Circumstances simply conspired to keep

[85] For the reasons familiar from discussion of such cases in epistemology, starting with Goldman 1976.

this from affecting much of what he got round to believing for a while. I submit that it is the underlying disposition that matters for interpretation; from this perspective, the early prevalence of fake barns, and Rogier's initial failure to recognise them, are mere noise. Simple knowledge maximization fails here because it is attuned to the noise.

As with the swan-case, we might consider the possibility that what matters is maximizing knowledge over time, but that move is vulnerable to the same objections as before. There is always the tragic Rogier, shot for trespassing at the end of the first day, upon, as he thought, attempting to examine a barnia more closely. Surely the ordinary and tragic Rogiers meant the same thing by 'barnia', given that the led the same lives and exemplified the same cognitive structures up to the fateful choice to trespass. Cumulatively, I take it that these two cases strongly support the claim that dispositions matter to interpretation. Insofar as knowledge maximization is not able to take dispositions into account, it is a flawed theory of interpretation.

Williamson provides an attractive theory of content that captures various intuitions about meaning. It also provides the clear epistemological basis lacking from the rationality-maximization theory. Specifying that interpretations map entire cognitive systems, including concepts that can be indefinitely combined, to contents clarifies how attributions of ignorance can be made, while appealing to naturalness can provide verdicts in cases which look like ties. However, this appeal to naturalness ought to be developed further to show exactly how it can handle more delicate cases. Finally, I argue that knowledge maximization as it stands at present is insufficiently sensitive to dispositions. A rationality-based theory might be able to handle dispositions better, so ideally we would like to find a theory that combines elements from both Williamson and Williams, uniting the strengths of both and balancing the weaknesses of each.

Chapter 4: Optimizing Dispositions to Know

After chapters 2 and 3, we wanted a way to combine the advantages of both knowledge and rationality maximization. Happily, Williamson describes a knowledge-first version of rationality: being disposed to conform to the knowledge norm of belief. Elaborating on this idea, I suggest that a theory of content on which the correct interpretation of an agent is that on which they are best disposed to know. I show how this proposal can handle the difficult cases where, in turn, rationality and knowledge maximization struggle by themselves. Finally, I argue that this proposal is a principled application of the knowledge first approach in philosophy, and give more general reasons for thinking that content is determined by an epistemic Principle of Charity.

I: Rationality and Dispositional Norms of Belief

In the preceding chapters, we examined two theories of content that fall under the broad umbrella of epistemic charity. Robert Williams advances a theory based on rationality; Timothy Williamson advances a theory based on knowledge. Both theories were found to be flawed. Williams requires a fully developed account of rationality that is able to handle the kind of cases to which externalists appeal. Williamson, meanwhile, should be more sensitive to dispositions. What we want, therefore, is a golden mean between the two approaches, combining their strengths in a way that will overcome their individual flaws. How can we have the subtlety and flexibility of a rationality-based approach, while retaining the epistemological advantages of a knowledge-first theory? By inserting, if we can, a knowledge-first account of rationality into the framework Williams provides.

Fortunately, Williamson himself provides just such an account of rationality.[86] On his view, the categories of rationality and justification are not central to epistemology: knowledge is what comes first. Accordingly, he is happy to recognise that there are different epistemic standards and ideals which interest us, various of which get loosely tagged with labels such as ‘rationality’ and ‘justification’. Analyzing rationality, as such, is not an important theoretical task; rather, what matters is that we differentiate the important epistemic standards and relate them precisely back to the central phenomenon of knowledge. Williamson believes that the most important epistemic standard is following the knowledge norm of belief: believe that P only if you know that P. But evidently there are other standards of interest to us, such as whatever standard we agree is met when we say that subjects are justified in Gettier cases. Such standards, Williamson argues, can be derived from the more basic knowledge norm of belief.

Where we value a norm, the thought goes, we also value the disposition to act in conformity with the norm. Take promise-keeping. Sometimes circumstances render a person incapable of keeping their promises, for instance, when they promised to be in London at 5pm but their 2pm train is delayed by two hours. If that person made reasonable preparations for reaching London by 5pm, such as booking a 2pm train for London and arriving at the station in time, then they are considered praiseworthy despite not strictly keeping their promise. This is because they evinced a disposition to keep promises, by taking actions beforehand which made fulfillment of their promise more likely. In most circumstances, those actions would have issued in a promise kept, and a person who habitually takes such actions will keep many of their promises. Conversely,

[86] In Williamson 2015.

they could have promised to be in London, attempted to board a train to Manchester, and then ended up in London on time anyway because they misread the timetable. Strictly speaking, the promise has been kept, but the actions taken did not evince good promise-keeping dispositions. In most circumstances, trying to take a train going in the opposite direction from the place one promised to be will lead to a broken promise. In general, for any norm N, we can derive a secondary norm DN: act according to the disposition to conform with N. The two norms are intimately related, but unlikely to be equivalent: for most norms, either can be complied with independently of the other. Both will matter for our normative evaluations of acts and agents.

This applies equally to norms of belief. Whether we conform with the knowledge norm in a given case is one question. Whether we are disposed to conform to the norm is another. If I check the time by a clock that happens to be slow, I will not comply with the knowledge norm of belief, but I followed good dispositions, carefully examining what evidence I had about the current time. In most circumstances, such actions would have led me to comply with the knowledge norm. I would have looked at an accurate clock and thus come to know the time. For that reason, I am somewhat praiseworthy. Hence Williamson identifies two related norms for belief. There is the primary knowledge norm, or KN: believe only if one knows. Derivative upon this, though, is a secondary norm, DKN: follow dispositions to believe only what one knows. When I follow good epistemic dispositions but circumstances conspire to keep more from knowledge, I am doing well by an important epistemic standard. It is this secondary and derivative standard, Williamson urges, which we have in mind when ascribing justification in Gettier cases and sceptical scenarios. A related term we might use for it, of course, is ‘rationality’.

This conception of rationality, I believe, is the starting point for a successful theory of interpretation. It is firmly grounded in an externalist, knowledge-centric epistemology, but also dispositional. We are not so much interested in specific beliefs, but rather the agent, and especially their cognitive system, as a whole. So the question is not how many beliefs conform to DKN and by how much, but simply how well disposed the agent is to conform to the norm of belief, or simply to know: the more belief-states that an agent would form are interpreted as knowledge, the fewer are interpreted as violations of the norm of belief. The correct interpretation of an agent is that on which they are best disposed to know.

We distinguish between the way things normally go, and the abnormal exceptions. Normally, salt dissolves in warm water, though this may fail in abnormal cases, such as when the water has already been saturated with salt. Normally, a bus will reach Roundhay Park from Leeds city centre in 15 minutes, though this may fail in abnormal cases, such as when there is heavy traffic. Normally, tapping an icon on my phone will initiate its respective app, though this may fail in abnormal cases, such as when the screen has frozen. Normally, my phone will accurately report the time, though again this may fail if the screen freezes. Normally, if I use my phone to check, I will come to know the time, but now you know the caveat. How exactly we should spell out this notion of normalcy is a matter for debate, but it seems to have more than a straightforwardly statistical force. We might describe a family as *normally* gathering for Christmas each year, even if a close look at the record shows that in most years, some novel, unanticipated circumstance prevented this from happening: a pandemic here, a snowstorm or an injury there.

One way to gloss this is in terms of explanatory priority.[87] The family's gathering at Christmas is the default case, requiring no special explanation. The frequent departures from this default each demand their own explanation, be it pandemic, snowstorm, or injury. Likewise, my coming to know the time when I check my phone is the default case, and whenever I fail to learn the time this

[87] See Smith 2010.

way some explanation is required, such as the fact that the screen has frozen at an earlier point. Normal occurrences themselves may or not have specific explanations: the family gathers because its members want to share Christmas together, and I learn the time because my phone displays it accurately. A sequence of coin tosses, however, will not generally be susceptible of explanation: it transpired through chance. The overall assumption is that there is an *explanatory order* to things, and many events, chancy or otherwise, will fall within that explanatory order, while others will not. In the case of a coin, a strictly equiprobable sequence of only heads would suggest a departure from the explanatory order of fair coin tossing, and the intervention of another explanatory mechanism, such as coin loading. Normal events fall within the explanatory order, abnormal events depart from it.

Many consider normality important for epistemology. Smith explains justification in terms of it, and Williamson uses it to gloss DKN. Our beliefs ought to follow dispositions that yield knowledge specifically in *normal cases*. The case where my phone displays 9:30, and the time is in fact 9:30, is a normal case. It is within the explanatory order of things for my phone to display the correct time. If it is in fact 10:30, the explanatory order has been disrupted and some special explanation is required, such as the failure of the phone's internal mechanisms, or my own cognitive functions. If I see 9:30 displayed and believe that it is 10:30, then I have followed a disposition that will lead me away from knowledge in normal cases, where the displayed time is correct. Thus I have followed a bad epistemic disposition despite believing truly. If, on the other hand, the phone's clock has failed and I form the mistaken belief that it is 9:30, then even in my error I follow a disposition that leads to knowledge in the normal cases where the time is as the phone displays and the explanatory order is undisturbed. Believing my phone evinces a better epistemic disposition, because it yields knowledge in normal cases.

Some have complained about this normalising turn in epistemology[88]. The basic worry is that normality is an intuitive notion on which we have too weak a grasp, and thus unfit to do the theoretical work to which it is put. According to Smith, for instance, winning the lottery is normal: that's why, on his view, if you falsely believe that you'll lose the lottery, your belief is not justified. By contrast, lies and stopped clocks are supposed to be abnormal, which is why beliefs formed on the basis of either lies or stopped clocks can be justified. But surely there is a perfectly good sense in which lies and stopped clocks are normal: at least as normal, we might think, as winning the lottery. So this invocation of normality to motivate a contentious epistemological claim is dubious; more generally, normality is not an adequate foundation for serious epistemological theorising.

This objection is overly hasty. In fact, the contrast between statistical probability and explanatory normality gives us a firm grip on the way in which winning a lottery is normal and lies or stopped clocks are not. How are the outcomes of lotteries explained? Some mechanism (say, the drawing of balls from a bowl) randomly selects a winner from among many entrants. Because the number of entrants is large, the probability of any given entrant winning is small, but equally small in each case. That one such winner is selected is certainly normal: in some, but not all, ways of administering a lottery, the failure to select a winner would in fact be abnormal. Why should it be abnormal that I am selected as winner? Well, given how lotteries work, *no reason at all*. The selection of any given entrant as winner falls entirely within the explanatory order of lotteries. To suppose that there is something special about *my* winning, such that it would be an abnormal outcome for a lottery, would be sheer egocentrism. Compare the family gathering at Christmas: whenever it happens, each party arrives at the gathering point at some specific time rather than another. That all the parties arrive at the exact times that they do, rather than any others, is very improbable. Yet that sequence of timed arrivals falls

[88] See McGlynn 2012b, Anderson 2017.

within the explanatory order of the family Christmas, as many other roughly equiprobable sequences would.

How, on the other hand, do lies work? To lie is knowingly to make a false assertion. Assertion is a practice whose purpose is to share knowledge, or at least truths; it is governed by a constitutive norm which lies violate.[89] Lies are parasitic upon sincere, norm-fulfilling assertions: if no norm-governed practice of assertion existed, then lies would not effectively disseminate false belief. Of course, people do lie, and quite often: but once we understand the position of lying within the context of the social practice of assertion, it is easy to see how lies falls outside of the explanatory order of assertion. Sincere assertion is the default; lies are a departure demanding explanation. Clocks too are precisely constructed to tell the time accurately; by means of intricate internal mechanisms they consistently display the correct time. That is how clocks work. It does happen that those mechanisms sustain damage, but such damage is what explains a departure from the default for a clock, which is to tell the time aright. Once we think through the explanatory order of each domain, we can see clearly what is wrong with McGlynn and Anderson's intuitions that winning the lottery is no less abnormal than lies or stopped clocks. The former is within the explanatory order of lotteries, the latter depart from their respective explanatory orders. In what follows, when I make judgements about normality in potentially contentious cases, I will not simply rely on bare normality-intuition, but make argument about explanatory orders such as those given above. Perhaps I may err, and no doubt the ideology of explanatory order is less than perfectly perspicuous, but we need not be as pessimistic about normality as are Anderson and McGlynn.

The position that I have outlined is similar to the explanatory link account of normality given by Andrew Lavinn.[90] Lavinn discusses the general phenomenon

[89] See Williamson 2000, Chapter 11 for a defence of the view that knowledge is the constitutive norm of assertion.
[90] Lavinn 2019.

of normality, without any specific reference to epistemic cases. This is in addition to a wealth of recent work on normality and epistemology specifically.[91] My use of normality is doubly applied. Not only am I applying a general account of normality to epistemic cases, but I drawing on the extant resources of normalised epistemology to theorise about content. The correct interpretation of an agent, on my view, is one which optimizes their epistemic dispositions. Let us return to the case where my phone incorrectly displays that is 9:30 am. Given two different interpretations of my mental states, therefore, one on which I believe truly that it is 10:30, and another on which I believe, following the disposition to trust my phone's display, that it is 9:30, the latter interpretation is correct. Though I am equally ignorant in either case (I presently have no way of knowing that is 10:30, the truth of my belief notwithstanding), I follow the better disposition if I believe that is 9:30. The correct interpretation of an agent is that which maximizes their knowledge across normal cases.

II: The Hard Cases for Williams

Earlier, we considered four cases for which we might fear that rationality maximization is inadequate (2.IV). *Lucky Brain* I will pass over for now, since it involves an envatted brain, and a full discussion of such cases will be given in the next chapter. *Modest Memory* asked why, given that it is rational to follow the apparent promptings of memory, rationality maximization could pick out a unique referent from a sparse internal description such as 'He had red hair and always talked about his squash team'. Whatever the reference of 'he', the subject believes rationally because of the rationality of following the prompt. This was not supposed to be a great challenge: any account of rationality beyond the most basic, seemings-based internalism should cope. In any event, it is clear that acquiring beliefs about random strangers whenever a memory passes through

[91] See eg Goodman and Salow 2023; Carter and Goldstein 2021.

one's mind will tend to lead one away from conformity with the knowledge norm of belief. Given that Sally is retrieving her past experience of one specific squash playing red-head, she is disposed to know with her use of 'he' if it refers to that man, and not otherwise. That is why optimizing dispositions to know yields the right result.

In *That Many*, it was supposed to be rational for Hardy to believe that there are 21 prime numbers between 1 and 100, even though he was in no position to know this. But the mere rationality of this belief was not enough to influence Hardy's content, which supported knowledge-maximization against rationality maximization. This case has a very simple solution on my theory: the specific norm of belief this case trades on - believe what is probable on one's evidence - may be one of many norms of belief for which 'rationality' is sometimes used, but it is not the norm that determines content. The relevant norm is *follow dispositions to believe only what one knows*. Evidently it would not be rational in the sense of conforming to *this* norm were Hardy to believe that there are 21 prime numbers between 1 and 100. We have already agreed that this is not something Hardy is in a position to know. More generally, believing arbitrary mathematical hypotheses which one is not competent to evaluate is a bad disposition, even if it sometimes issues in true beliefs, shy of knowledge. So the optimizing dispositions to know view makes no implausible prediction that Hardy's use of 'that many' refers to 21.

A more interesting case is *Red Bag*, based around Maria Lassonen-Aarnio's argument that knowledge and dispositions to know sometimes come apart from one another[92]. If Suzy believes things are as they appeared to her perceptually, then she knows that the bag in the middle was red, or so knowledge-first epistemologists say. But if she believes this, then she ignored the defeater provided by John's misleading testimony, and so followed an underlying disposition that tends away from conformity to the knowledge norm of belief,

[92] Lasonen-Aarnio 2010

even if it gained her knowledge in this one unusual case. Since I argued that the correct interpretation is that on which she refers to the bag that seemed, and was actually, red, it looks like simple knowledge maximization fares better than the dispositional account here.

However, I mentioned previously that the reason why this interpretation is so compelling is that Suzy's irrational ignoring of testimony is already built into the scenario as described. Suzy failed to reflect on relevant evidence. This is irrational, but that irrationality is a fixed feature of Suzy's behavior, and not an artifact of any interpretation. To put it in the terms we are now using, it is part of the story that Suzy evinces a disposition to ignore testimony that tends away from conformity to the norm of belief. This bad disposition remains whether she refers to the red bag or the blue. However, if she refers to the red bag, she is at least following one good disposition: that of believing what her perceptual evidence supports. If she refers to the blue bag, she is following two bad dispositions: the original disposition to ignore testimony, and also a further disposition to believe contrary to the evidence of her perception. She is better disposed to conform to the norm of belief if she has one good disposition than two bad ones, and so optimizing dispositions to know gets the right result after all.

III: The Hard Cases for Williamson

In McGlynn's *Hallucinogen* case (3.IV), Lucy seemed to see several people in front of her. Some of the person-seemings were caused by Helen, a woman standing in front of her. She was the only person in front of her, and the rest of the person-seemings were caused by the hallucinogen. When Lucy formed beliefs while attending to Helen, she refers to Helen, but because of her hallucinations those beliefs were unsafe, and thus not knowledge. How, on my view, is this reference secured despite Lucy's ignorance? My starting point is that, insofar as Lucy is hallucinating, this is not a normal case. In the default case, when Sally seems to see a person, she is receiving light which has reflected from a person. Such is the explanatory order into which person-seemings fit. The extra person-seemings in *Hallucinogen,* not caused by the reflection on light from people, cry out for a special explanation, which Lucy's ingestion of the hallucinogen supplies. In normal cases, when Lucy attends to Helen and forms the belief expressed by 'she is freckled' without any accompanying hallucinations, Lucy expresses knowledge if 'she' refers to Helen. Thus on the interpretation that optimizes Lucy's dispositions to know, 'she' refers to Helen.

This said, I improvised upon *Hallucinogen* to provide a further challenge to knowledge maximization, which I had better rise to myself, though I shall be discussing similar issues in more detail in the next chapter. In the *Hyper-Hallucinogen* case, Lucy takes a slightly different drug, which has the effect of screening-off the vision centre of Lucy's brain from her eyes entirely. She still seems to see several people, including a tall, freckled etc woman, but Helen does not cause any of these seemings. In this case, Lucy fails to refer to Helen. So the task is to specify why Lucy fails here when she succeeds in the original *Hallucinogen*. Once again, we may observe that this is an abnormal case. The question is, how do the dispositions Lucy follows here fare in normal cases? The relevant disposition is presumably attending to qualitatively identical, or at least relevantly similar, person-seemings to those Lucy undergoes in *Hyper-Hallucinogen*. In normal cases, attending to such person-seeings would yield knowledge of a woman causing Lucy's person-seemings. But nothing that we have said suggests that they would uniquely yield knowledge of *Helen*, the specific woman in front of Lucy during *Hyper-Hallucinogen*. Any woman who can cause sufficiently similar seemings will do. Since normality rather than closeness is our concern, the proximity of Helen in this case is moot. Since there is no causal relationship between Helen and Lucy's seemings, we would not expect any close resemblance between them. Even if we write an accidentally close resemblance into the case, this would not rule out Lucy undergoing a perfectly normal encounter with a very similar looking woman to the Helen of *Hyper-Hallucinogen*. Therefore the interpretation on which Lucy's use of 'she' refers to Helen in *Hyper-Hallucinogen* does not uniquely optimize Lucy's dispositions.

I then introduced two cases of my own (3.VII). Beatrice heard reports of Australian fauna halfway through her life, thus changing her attitude towards the English sentence 'All swans are white' from assent to dissent. We suspected that Beatrice has a single swan-concept with a stable content throughout her life, and always referred to the unknown Australian swans. Simple knowledge maximization struggles to predict this, though, as she knows more before the testimony comes through if she refers only to European swans. Our new account does better. Suppose that she initially refers only to European swans. Then her 'All swans are white' belief is a product of careful induction from ample evidence for a fairly restricted kind, successfully yielding knowledge in this instance. She is following a good epistemic disposition. However, she also has latent within her a much worse disposition: the disposition to believe that *European swans* are black on the basis of reports from Australia. This runs clean contrary to the knowledge norm of belief, as is evident when the disposition manifests. On the other hand, if she refers to swans in general, then she is still following a fairly good inductive disposition, albeit given the generality of the concept one more vulnerable to exactly the sort of surprise counter-example as actually afflicts it. However, this is more than compensated for by her disposition to respond to such discoveries precisely as counter-examples. Given that the concept is general, this disposition tends to ensure conformity with the knowledge norm of belief. Overall, then, the European swan interpretation leaves Beatrice with one good disposition and one bad disposition, while the generic swan interpretation leaves her with one good disposition and one fairly good disposition. So the generic swan interpretation is correct, since it leaves Beatrice better disposed to know.

Next, Rogier introduced the term ‘barnier’ to describe the structures he sees in fake barn country. At first, he applies it on seeing fake barns, then he starts to see more real barns, and eventually he learns about the presence of the fake barns and tries to reserve ‘barnier’ for the real barns. So what did the term mean when he had only seen fakes? Suppose that ‘barnier’ refers indifferently to fake and real barns in the relevant style. Then Rogier followed a good disposition in saying ‘There’s a barnier’ as he did on the first day: he was assessing his perceptual evidence well, consistently gaining knowledge as a result. But, again, within him was latent that bad disposition to say of known fake barns that they are not barniers. On the indifference interpretation, this leads away from knowledge. On the real barn interpretation, Rogier’s initial disposition is still quite good: he is assessing his perceptual evidence in a way that would yield knowledge in most cases, but unusual circumstances thwart him. Meanwhile, his latent disposition to deny that known fakes are barniers is vindicated: it tends strongly towards conformity with the knowledge norm of belief. So the final verdict is the same as for Beatrice: on the knowledge maximizing interpretation, Rogier has one good and one bad disposition. On the interpretation where early and late Rogier’s content is consistent, he has two good dispositions from the beginning. So the consistent interpretation is that on which Rogier is always best disposed to know, and thus correct according to my theory.

IV: A Principled Principle

We have found an account of interpretation that will fill the general niche we had identified, and solves the cases that had plagued other theories. Ideally, however, we would like a theory that does more than just satisfy a few intuitions.

Proceeding in such a manner courts overfitting: even granting that Williamson's response to judgement scepticism is successful, we are not infallible arbiters of hypothetical cases.[93] I may have erred in evaluating cases; insofar as my theory has been tailored to explain a judgement that is in fact erroneous, it will be flawed. As more cases unreviewed here come to light, then this error-driven theory will struggle to explain them, forcing revisions designed to accommodate all of our judgements about the new cases, some of which may in turn be erroneous. As this process iterates, our theories of content become more complicated, and the search for a plausible theory of interpretation looks increasingly like a degenerating research programme.[94] To supplement argumentation from cases, therefore, it would be desirable to have more direct arguments showing that a given theory is supported by deeper principles.

Let's start from the assumption that representations aims at truth.[95] A belief is a means of representation, and so, broadly speaking, beliefs are directed towards truth. But beliefs are stable representations; commensurate with this stability of representation, they aim more specifically at *stable truth*: knowledge. At a higher level, a cognitive system is a means of representation, and so, broadly speaking, cognitive systems are directed towards truth. But cognitive systems enable representation by generating beliefs, which aim more specifically at knowledge. Furthermore, they are stable systems for generating beliefs; commensurate with this stability of belief-generation, they aim more specifically at the stable generation of knowledge: optimal dispositions to know. Below I fill out this argumentative sketch, discussing along the way some alleged counter-examples to the knowledge aim of belief.

[93] See Forster and Sober 1994; Williamson (forthcoming).

[94] See Lakatos 1970. Forster and Sober argue that what's wrong with degenerating research programmes, and the absence of theoretical simplicity more generally, is in fact to be explained in terms of overfitting. The simpler the curve used to fit data, the more predictively accurate the theory will be, since its predictions are less likely to be influenced by past errors in the data.

[95] The argument offered here is inspired by the account of perception in Burge 1994, as well as the ideas in Nagel 2023. Beyond this, there is an extensive battery of arguments for the conclusion that knowledge aims at belief: Bird 2007 and McHugh 2011 are fine defences of the view.

Imagine a creature with a simple system of perceptual representation: it continuously receives light from its immediate environment, and in processing this light it generates a stream of momentary representations of its immediate environment. In order to guide the creature successfully through the environment, these representations should be accurate: their goal, we may assume, is truth. There is a log before it, and looking at the log the creature represents that there is a log before it, moment by moment, until it focuses its perceptual attention elsewhere. Because there is a log before the creature, each momentary perceptual representation that there is a log before the creature fulfills its goal. This stream of accurate representations allows the creature to move successfully around the log. All is well with the creature, its representational system, and the momentary representations that the system generates.

Notice, however, that this happy case for the creature requires a great deal more than the accuracy, and so the goal-fulfillment, of any specific momentary representation. At a minimum, these representations must be consistently accurate, at least for as long as the creature is using them to navigate the log. But even this is not as good as it can get for the representational system. If the system could easily have produced some or many false representations, then the system would be vulnerable or fragile in some important respect, given its goal of producing accurate representations. It would be doing badly. It would be doing better if it could not easily produce false representations, if its output of accurate representations were safe. So while the momentary representational system may have no more demanding goal than truth, it's plausible that the representational system overall has the further goal of safely generating accurate representations.

Let us now consider the case of ordinary human belief. Our beliefs are not like the momentary representations discussed above; they are fixed representations. The creature navigating its way around a log on the basis of a stream of momentary perceptual representations of a log is broadly comparable to navigating your way around a log on the basis of the belief, say, that there is a log 5 paces long 3 paces in front of you. It's like the creature taking a single one of its momentary perceptual representations, or (more realistically) a model extrapolated from enough of its momentary perceptual representations, and storing it in memory. Your action is not being guided by a series of many representations, but by a single representation. A belief, we might well think, is a system for ensuring that we represent the content of that belief for as long as we hold the belief.

Merely representing that P - in this case, a relevant proposition about the log - is entirely proper given the truth of P, per our initial assumption, and this so regardless of whether we represent that P through believing that P or having a perceptual representation that P, as the creature does. But the propriety of believing that P - that is, storing a fixed representation that P - is another matter. It is subject to the same sorts of consideration that we have seen apply to the creature's simple perceptual system. The belief is in a bad way if it is sometimes false and sometimes true, and it is still in a bad way if it happens to be consistently true, but could easily have been false. Its goal as a representational mechanism is only fully served if it safely stores accurate representations, just as the simple perceptual system's goal is fully served only if it safely generates accurate representations. Therefore knowledge is the goal of belief.

Another way to think of it: when we believe that P, we stably represent that P. Not come what may, exactly, since we do in fact change our beliefs over time, but come much that may to which we have no direct access. In a close case, you would expect to represent a log perceptually, maybe much more finely than the creature can. But that is not so with, say, your belief that Ottawa is the capital of Canada: if the Canadian government were to discuss changing their capital, you probably wouldn't be perceptually representing it. This stability in representation ought therefore to be accompanied by a commensurate stability in the accuracy of representation, given the goal of accurate representation. Not, again, come what may, because we do change our beliefs, but come much that may to which we have no direct access, and which accordingly would not impinge upon our beliefs. Storing a representation that Ottawa is the capital of Canada, by believing that it is, only makes sense to the extent that this representation is safely accurate. So, again, knowledge is the goal of belief.

It has been argued that in some cases, belief without knowledge can be entirely proper, or fully justified, and so that knowledge is not the aim of belief. Two examples are discussed by McGlynn.[96] In one, Jane believes that her lottery ticket will lose. The odds against her winning being very high, this belief is justified, and in this instance true. Yet, per the dominant assumption among epistemologists, and more specifically because of the failure of safety, this belief is not knowledge. Yet McGlynn claims that this is a fully justified belief, and indeed Jane herself can reasonably maintain this belief even acknowledging that it fails to constitute knowledge. In the second, Captain Jack Aubrey has extensive experience of French naval strategy, and having monitored the manoeuvres of the French fleet all morning, comes to believe that they will attack after nightfall. This is a good hunch, based on expert knowledge and careful evaluation of the evidence at hand, but not sufficiently secure as to constitute knowledge. Yet McGlynn argues that this is again a fully justified belief, and again that Aubrey might reasonably retain his belief even accepting that it is not knowledge.

I think trying to overturn the knowledge account of belief on the basis of such examples would indeed be overfitting. First, some considerations against the full justification of these beliefs. In both cases, there are alternative beliefs available: Jane might believe that her ticket will *probably* lose, and Jack that the French will *probably* attack after nightfall. McGlynn denies that they believe (no more than) these alternatives in the relevant cases, as he may fairly stipulate. My point is no more than that these alternatives might have been settled for. So, why

[96] McGlynn 2023; see also 2011.

accept the stronger beliefs? Plausibly, beliefs should be proportioned to the available evidence; plausibly, in these cases, the subjects fail to proportion their beliefs to the evidence. Believing the weaker claims would be proportioning belief to evidence; in believing the stronger claim, they are exceeding their evidence. Indeed, McGlynn goes so far as to specify that Aubrey's evidence is inconclusive.[97]

Of course, one might likewise object to the belief that the French will probably attack after nightfall: why not simply hold the even weaker belief that it is probable that they will probably attack at nightfall? On pain of regress, this reasoning cannot be extended indefinitely. I think the best stopping point is knowledge: it makes sense to ask, when Jane has inferred from the known probabilities, or Jack has ceased surveying his inconclusive evidence, why each does not settle for the weaker belief. It is then a good question. It would not make sense to ask this once the results have been reported, or when the French are attacking at night: once knowledge is available, it becomes an idle question. Thus in both of McLens cases his subjects are not justified in holding that beliefs rather than some weaker known alternative they do not show they failed to show that knowledge is the aim of belief

More than this, we can explain McGlynn's judgment to the contrary. Returning to Williamson's promise keeping analogy, suppose you have promised to meet your friend in the cafe at 1:00 p.m. but you arrive at 1 minute past. It would be entirely inappropriate, in fact vicious, for your friend to criticize you. You were only one minute late, and they have been sitting comfortably in their armchair in a fine cafe. Nonetheless, you failed to keep your promise, and violated the norm of promise-keeping. The violation was simply trivial, and pointing it out would be carping. This, I submit, is what is happening in McGlynn's cases. Both beliefs are close to knowledge: close enough, we might think, that no practical difference is made. Jane's expected utility from holding on to the ticket is tiny,

[97] McGlynn 2023, p 20.

and she might well keep it and enjoy watching the draw anyway; meanwhile, Jack cannot avoid acting under uncertainty. These are trivial violations of the normal belief, and actually to criticize them in a real social social situation would be inappropriate carping. This explains our reluctance to find any fault in Jack or Jane, even in the theoretical context. Nonetheless, they are at fault: they have slightly violated the norm of belief, just as you slightly violated the norm of promise-keeping by arriving a minute late.

Having defended the knowledge aim of belief, let us return to the main thread of our argument. Recall that we started out with a truth aim for representation, and concluded that, since beliefs ensure stable representations, they should aim at truth in a correspondingly stable way. The next step is to argue that a cognitive system as a whole stands to belief much as belief stands to representation as such. Cognitive systems stably generate beliefs, just as beliefs stably represent propositions. We combine and recombine our contents in predictable ways to generate new beliefs in new situations: not come what may, exactly, since we can revise our concepts and change our patterns of belief formation, but come much that may. Updating our beliefs about the natural history and internal structure of whales after investigation is one thing; reconceptualising whales as mammals rather than fish is a longer and more laborious process. Accordingly, our cognitive systems should generate knowledge come much that may. They should not be vulnerable to exigencies of circumstance, such as passing through fake barn country, or being separated from subspecies. The commensurate stability in knowledge generation, to accompany the stability in belief generation, is a matter of optimizing dispositions to know.

Consider a cognitive system as a tool for generating knowledge, and compare it with other tools. A person uses their cognitive system to gain knowledge, just as a musician uses a cello to produce music. It might be that a cognitive system does better the more the person whose cognitive system it is knows. But it is not only the cognitive system that contributes to knowledge: the person must deploy their concepts well, and the world must co-operate. An excellent cello can fail given the goal of making good music because I am playing it, or because it is being drowned out by a loud and poorly tuned orchestra. Likewise, a mediocre cello can produce good music because a superb cellist is playing it accompanied by a good orchestra. The contribution of the cello itself to the goal can be evaluated independently of how well the goal has been realized overall.

The cello contributes by enabling the player to make good music. The better the music made, given the same actions on the player's part, the better the cello. At least, the better the music made *in general*, the better the cello. The music may be better or worse because of the accompaniment, and that is no contribution of the cello's. So what matters is the player's *underlying disposition* to make good music. The further the cello enhances that disposition, the better it is. Just as the cello is a tool for making music, so a cognitive system is a tool for gaining knowledge. A cognitive system is a good tool to the extent that it enhances a person's general disposition to know. We initially set out to find a Principle of Charity. We might think of the general form of a Principle of Charity along these lines:

> The correct interpretation of *x* is that according to which *x* does best given goal G.

Taking knowledge as our goal G, and cognitive systems as our direct objects of interpretation *x*, we obtain the following principle:

> The correct interpretation of a cognitive system is that according to which the cognitive system does best given the goal of knowledge.

As argued, a cognitive system does best given the goal of knowledge to the extent that it optimizes a person's dispositions to know. So we have a direct argument for optimizing dispositions to know, proceeding from deeper theoretical motivations, rather than a mere survey of cases. Optimizing dispositions to know is a principled principle. As we have seen, however, it is also supported by a survey of cases, combining the strengths of both knowledge and rationality maximization, while successfully handling a range of scenarios that these alternatives struggle with on their own. Thus I suggest that the correct interpretation of an agent is that on which they are best disposed to know.

Chapter 5: Exploring the Edges

Now that we have a theory of content in hand, and considered it in relation to some specific cases that trouble rivals, we can apply it to some big picture questions about interpretation. Putting the theory to work allows us to flesh out the finer details, as well as evaluate its theoretical fruits. It may even shed some light on some independently interesting topics. The two basic questions I want to consider are how we interpret an agent who lacks knowledge, and how we interpret an agent who lacks history. Accordingly, I will discuss envatted brains, then swampmen, before trying to combine the two kinds of case.

A: Envatted Brains

I: Introduction

BIVs are an intriguing test case for any Charity-based theory of content, since it is unclear how charitable one can and should be to an envatted brain. More alethic theories will generate interpretations easily: just feed your agent the facts, whatever they may be. But this is exactly the kind of case which rouses suspicion about such theories: we naturally assume that the envatted brain is deluded. More epistemic theories will avoid attributing implausibly accurate beliefs, but face a problem of their own. If positive epistemic status is the basis of interpretation, how does one go about interpreting an agent in so poor an epistemic position? A comprehensive theory of content will have something to say here, especially given the wider philosophical interest of the question: the beliefs of the BIV are crucial to discussions of scepticism.

‘You may think that you have hands’, the student is warned, ‘but so does an envatted brain’. The naked brain, kept alive within a vat and subjected to electrical stimulation which replicates the experience of embodied interaction with the external world, is a scenario by which many are introduced to the issue of scepticism. There is a natural reading of this scenario. Just as the brain has experiences like ours, though simulated, it also has beliefs like ours, though false. The brain wrongly thinks that it has hands, that Sally is running her hands through her hair, that swans are gliding on the lake, and so on. This would be an *error theory* for envatted brains: BIVs have many beliefs about an external world independent of its experience, but these beliefs are mostly false.

While the error theory may be natural, it is by no means an obligatory interpretation of envatted brains. Famously, Hilary Putnam rejected in his 1981 paper, citing a causal constraint on content.[98] His work is part of a broader anti-realist tradition of using specific accounts of content to exclude the possibility of radical error.[99] Such strategies, however, are not universally appealing: obviously they are not congenial to sceptics, but they are also dissatisfying to *anti-sceptical realists*. The anti-sceptical realist wants to maintain that radical error is a genuine possibility which we have avoided, and thus that we are *better off* than the benighted brains in vats. Both the sceptic and the anti-sceptical realist will prefer a theory of content that supports the more intuitive error theory.

There is a final alternative, however, at what we might think of as the opposite extreme from anti-realism. On this view, not only do envatted brains fail to hold true beliefs, they fail to hold many beliefs at all, true or false. Since content

[98] Putnam 1981.

[99] See, for instance, the overview in McKinsey’s 2018 discussion for the SEP.

requires a positive epistemic status - knowledge, let's say - and BIVs lack knowledge quite generally, they also generally lack any basis for content. The brain may seem to enjoy a complex cognitive life, but, like so many other aspects of the brain’s life, that is an illusion. The brain has no comprehensive system of beliefs comparable to our own either in structure or content. This is the reference failure interpretation of BIVs. I think it is suspicious, much as the opposite extreme is suspicious: it conflicts with our assumption that the envatted brain is deluded. Ideally, we would avoid both of these extremes, and retain the intuitive error theory.

In what follows, I will argue that my view of content supports such an error theory for BIVs. Since I take it that this is the most plausible interpretation of envatted brains, and suits well what many would already want to say about scepticism, I believe this to be a benefit of my view.

II: The Pitfalls of Simpler Knowledge Theories

In this thesis, we have been concerned with Charity as an approach to content. Roughly, what people believe is whatever it would make most sense for them to believe. Their beliefs may make sense because they are true, or because they are reasonable. There are different ways of explicating this idea, and many of them will spell trouble for the error theory of envatted brains. Most obviously, a principle of charity that straightforwardly maximizes truth will always exclude any possibility of radical error: albeit such a conception of charity is so crude as to have few defenders, and is mostly useful as a toy model. Other variants of interpretationism, however, raise similar issues. Take the family of principles based on rationality. On a sufficiently *externalist* account of rationality, on which it is a matter of being connected to the world in the right way, we will see a parallel preclusion of radical error. If, for instance, rationality is a matter of *tending to believe that p just when p*, the brain in a vat will tend to believe truths, and not erroneously imitate the beliefs of an ordinary human in an entirely different situation. Indeed, Donald Davidson, one of the most important figures in the development of interpretationism, advanced anti-sceptical arguments along precisely these lines: since truth is crucial to interpretation, the nature of content limits the possible scope of error.[100]

For this and other reasons, we might suspect that truth is too loose a standard for interpretation. Intriguingly, however, similar results threaten even an approach as strict as knowledge maximization. Since knowledge is factive, such a view will exert a similar pressure against radical error. An envatted brain does not know that it has hands, and so an interpretation of the brain on which it believes that it has hands will, per knowledge maximization, be defective in this respect. We might think, however, that the brain can know about hand images, or permanent possibilities of hand-sensation. It knows that it has two particularly easy such possibilities of hand sensation, direct control over which enables a measure of indirect control over its other sensations. In order to maximize the brain's knowledge, therefore, it looks like the knowledge maximizer is going to have to give the brain a phenomenalist interpretation, according to which the brain refers only to the object of its own experience, or perhaps objects constructed out of its experience.

What makes this result surprising is that knowledge maximization might be expected to pair with anti-sceptical realism: Timothy Williamson, for example, is an archetypal anti-sceptical realist.[101] So how might knowledge maximization be reconciled with anti-sceptical realism? In the first instance, one might deny that a thoroughgoing phenomenalist interpretation of the brain is even available. Such an interpretation, we might think, requires ontological resources that just aren't

[100] See, for instance, Davidson 1975.

[101] The last two chapters of Williamson 2007, including his defence of knowledge maxmization, function as an apology for anti-sceptical realism in metaphysics, as we saw in our examination of his theory.

there. It is one thing to talk of mapping the brain’s concepts onto objects of its experiences (or constructed out of its experience), but quite another actually to provide such a mapping. Secondly, one might argue that knowledge of its own experience is harder for the brain to acquire than we are apt to suppose. A central theme of Williamson 2000 is that our knowledge of our minds is not especially secure, to the extent that our knowledge of our immediate environment may even be better than our knowledge of our experience.[102] Together, these ontological and epistemological objections to the phenomenalist interpretation may prove powerful, and so a knowledge maximizer might argue that we ought instead to attribute massive reference failure to the envatted brain. On this view, the brain may seem to enjoy a complex cognitive life, but, like so many other aspects of the brain’s life, that is an illusion. The brain has no comprehensive system of beliefs comparable to our own either in structure or content.

While this may be a viable interpretation of the envatted brain that avoids attributing it with easy knowledge, I still contend that the error theory is more natural. The alleged gap between what beliefs the brain actually has and what it appears to have is enormous. The brain also presumably has higher-order beliefs about its own beliefs, which would in turn be radically mistaken. We need not be unreconstructed Cartesians to balk at such a failure of self-knowledge, especially if we are trying to maximize knowledge overall. Further, as we saw McGinn argue in 1.V above, the neural structure of the envatted brain is very similar to that found in agents with complex belief-systems.[103] Again, we need not be extreme internalists to be surprised that the cognitive differences could so vastly exceed the neurological differences between the two cases. It would be better to avoid the reference failure interpretation if we can, and seek an error-theoretic alternative.

[102] Williamson 2000. See eg the admonitions against interiorising evidence on p 193.

[103] McGinn 1986 p 186.

III: Dispositions and Error

My account of content can provide an error-theoretic alternative. Instead of maximizing knowledge, I aim to optimize dispositions to know. The correct interpretation of an agent as that on which they are best disposed to know. An agent is best disposed to know, in turn, when they know the most across normal cases. This helps here, because all BIV cases are abnormal. In the default case, what explains the perceptual experience as of a chair is the receipt of light from a chair through the eyes. The human eye evolved in tandem with the vision centre of the brain, the former to receive light and the latter to convert that light received into perceptual experience, to increase our ancestors' overall fitness. Hence there is a deep explanation for why these two things - perceptual experience as of a chair and light received from a chair - occur together. BIV cases depart from this explanatory order, and thus call out for special explanation. Only through the complex interventions of the scientist did the perceptual experience as of a chair come together with direct electronic stimulation within a vat. This means that, on my view, BIV cases are never relevant to interpretation, not even to the interpretation of a BIV. Instead, we interpret a BIV by trying to optimize their underlying epistemic dispositions, which evaluate according to their performance at worlds where they are embodied and interacting with their environment *as normal*.

The BIV who seems to see a chair will have what we might think of as a 'chair concept': a concept activated on seeming to see a chair, or indeed seeming to hear the English word 'chair'. The crucial question when it comes to assigning a content for this concept is how much knowledge the brain would encode with it, were the special circumstances interfering with its knowledge-gathering absent: that is, how much it knows by means of it when embodied. The brain actually uses the concept in response to chair and ‘chair’ seemings, and this pattern defines its underlying dispositions of use. The brain would use it when embodied in response to similar seemings. So it believes ‘There is a chair in front of me’ when seeing a physical chair in front of it, and ‘Dad sat on a chair’ when its father accurately reports sitting on a chair. Such beliefs will be knowledge if the concept refers to physical chairs. Thus such an interpretation optimizes the brain’s disposition to know, and though it remains ignorant within the vat, it does successfully refer to chairs outside the vat. The brain falsely believes that it has seen the physical chairs we embodied humans encounter.

As with chairs, so too with chickens and forests and bus drivers. The brain has concepts which it uses in such a way that, were it embodied, it would identify the object by means of the concept. Since such cases are the normal ones, it is to those objects that the concepts refer: that is how the brain is best disposed to know. Which means that the brain has all sorts of false beliefs about chickens and forests and bus drivers. The brain has a comprehensive system of beliefs about the external world, and those beliefs are radically mistaken. My view secures the intuitive error theory for BIVs.

IV: The Limits of Envatted Reference

At this point, we might worry about proving too much. The envatted brain is in a very bad way. We are much better off than it. Surely this should have some repercussions for the brain’s ability to refer? Consider proper names. Suppose the brain has been induced into a series of experiences as of a woman named Sally, where the Sally character is a complete fiction at most accidentally resembling any real woman. The brain will have a Sally concept, activated when it undergoes Sally experiences, using which it believes 'Sally is over there' and 'Sally is running her hands through her hair'. In such a case, the error theory does not seem enough, and a diagnosis of reference failure is more appropriate. A concept trained on illusory woman-experiences would not name a real woman. This is so even if, by chance, a real woman named Sally happened to stand a short way from the vat, running her hands through her hair: even one who (accidentally) resembled the brain's Sally-illusions. Its beliefs might be true, and even justified, if the concept referred to the woman outside the vat, but still the reference does not happen. Direct reference to specific things outside of the vat is beyond the brain's capacity.

In *The Philosophy of Philosophy*, Williamson uses such cases to defend his own view of content.[104] A strict knowledge-maximizing theory, he urges, best explains the failure of reference from within the vat. Recall the *Lucky Brain* case from 2.IV above. The brain does not refer to Sally because it does not know anything about her, the potential truth and justification of its beliefs notwithstanding.

Williamson does seem to make a fair point here. To assess the content of the brain's cognitive system, on my account, we need to assess what it knows using that system across normal cases. If the brain is using the same cognitive system, it will have a Sally concept that it uses in much the same way as its actual Sally concept. This in turn implies that the brain must have very similar experiences to the Sally-simulations it actually has, in response to which it deploys its Sally concepts. Such cases being normal, moreover, these experiences will be largely veridical. So there is a real Sally in this normal case who is exactly as the brain actually experiences the simulated Sally as being. And if the brain's Sally concept refers to this Sally, it will encode much knowledge across cases at which

[104] Williamson 2007, p 271. See the earlier discussion in (2.IV). esp. *Lucky Brain*.

it employs the cognitive system. So it looks like the reference of the brain's Sally concept ought to be Sally, the woman from the normal cases. Suppose further that the Sally actually outside the vat really is an exact, if accidental, match for the simulated Sally, and thus for normal case Sally. It now looks like one and the same Sally exists at the actual world, and at the distant worlds where the brain is in a position to know about her. If the brain's Sally concept refers to the actual Sally, therefore, it would know more across normal cases. So it seems that, on my theory, the correct interpretation of the brain is that on which its Sally concept refers to the real Sally that lives outside of its vat.

Nonetheless, the parallels between this case and the *Hyper-Hallucinogen* case I previously adapted from McGlynn (3.VI) should make us pause before awarding any trophies to Williamson. Veridical perception does indeed provide a basis for reference that mere accidentally accurate seemings cannot, but this is not always well explained by knowledge maximization, and indeed I already made a brief attempt to explain it in my own terms. I will elaborate on those ideas further here. For a start, I believe that need not presume reference to specific individuals outside the vat to be straightforwardly impossible. We need not presume that reference to specific individuals outside the vat is simply impossible from within. Consider a case where the complete (digitized) contents of the British Library have been uploaded into the simulation. The envatted brain undergoes an experience as of reading a biography of Wellington. This experience is based on an actual biography of Wellington, simulated verbatim. In this case, I suggest, the brain is able to refer to the historical Wellington, envatment notwithstanding.

Nonetheless, we might think that this example is consistent with a strict knowledge maximizing approach. Perhaps the brain knows less than we do, but if all the books it can 'read' are copied verbatim from the British Library, it could not easily have been mistaken about the details of Wellington's life. So substantial 'third person' knowledge of the external world is possible in British

Library BIV case, and it is this knowledge that establishes reference to individuals like Wellington.

The point is not so much to show that knowledge maximization goes wrong, as to illustrate that 'BIVs can't refer to individuals beyond the vat' is hyperbolic as an objection to my own view. Even if we allow Wellington, however, I owe some account of what the limits on BIV reference are. Let us return to Sally seemings. As an embodied human being, it is easy for me to start referring to Sally. If I have a fleeting Sally-glimpse, that is enough for me to use 'Sally' and mean Sally. When I say 'Sally just passed me by', I speak knowledgeably about Sally. I just saw her pass me by, after all.

So the worry is this. If I can refer to Sally after a glimpse, because when I glimpse her and gain knowledge, then a BIV can refer to Sally after a simulated glimpse, because if it glimpsed her the BIV *would* gain knowledge. Per my dispositional theory of content, this suffices for reference.

This worry can be alleviated by considering the crucial differences in the way that embodied human beings use their cognitive systems and the way that BIVs do. The BIV's use of 'Sally' is exhausted by responses to certain experiences. I may also use 'Sally' in response to experiences, but I also do more than that: I use it in response to *seeing Sally*. I see lots of things, Sally included, and I compare what I see to what I have seen of Sally. That is a crucial feature of how I use the name. The process, moreover, is already in full flow after one glimpse. Every subsequent glimpse can be compared to the first, allowing me to build up detailed knowledge of Sally. After one glimpse, there is a unique woman, Sally, about whom my use of the name encodes knowledge, and will go on to encode further knowledge,

The brain, boasting only a simulated glimpse of Sally, is in a completely different position, even granting that its experience is qualitatively no different.

Seeing Sally and comparing things to her is not an option for it. The physical proximity of Sally to the vat is moot: unlike in the normal case, the brain's Sally experiences have nothing to do with this proximity. So asking what the brain *would* know in normal cases does not just involve considering those cases where the brain sees the woman that is actually beside it in the vat. It involves considering any case at all where the embodied brain undergoes a similar but veridical experience. The brain could glimpse a completely different woman who happens to resemble Sally, and this too will be the brain applying its Sally-concept in a normal case. Because the only basis for the brain's usage is mere experience, and not a successful act of glimpsing, the concept is not 'trained' on a unique woman. Hence the brain is not, in using the concept, well-disposed to know about any specific woman, and so there is no specific woman to which the concept refers. A BIV cannot refer to an individual beyond the vat on the basis of a simulated glimpse, even if that individual is actually close by and similar to what the brain seemed to see.

This then is my account of the limitations on individual reference for BIV. For a concept to refer to a specific individual, it must be 'trained' on that individual such that in using it the agent is well-disposed to acquire knowledge of that individual. This 'training' of concepts is much harder from within a vat, as there are no direct perceptual links between individuals and the concepts. Even fleeting experiences can train a concept on an individual, if the experience is a veridical perception of that individual. If experience alone is all the agent has to go on, there is a much higher bar. Simulating an accurate, detailed biography might do, but seeming to see a rough likeness will not. Only a wealth of truly individuating details within the experience can compensate for the absence of a perceptual link.

V: Burge on BIVs

Before proceeding further, it might be useful to compare the views of Tyler Burge.[105] Burge takes similar ideas - error-theory for BIVs, normality, and epistemic norms - but proceeds in the opposite direction. For him, the entitlement to believe on the basis of perceptual experience is the *explanandum*, and successful reference to the physical environment is the *explanans*. Beliefs formed on the basis of certain perceptual states are warranted, because they, and the states on which they are based, are reliable in normal conditions; normal conditions are those 'by reference to which the nature of the perceptual state' or belief is explained.[106] Explaining the nature of a state, in turn, is, at least in large part, a matter of explaining how the state comes to be associated with its content: the normal conditions are those prevailing in, or relevantly similar to, 'the content-establishing environment'.[107]

Take, for instance, a perception as of a chair: this is the very perceptual state that it is, and not another, endowed with the content it has, rather than any other, because of complex interactions with the environment that enabled an agent to have some system that functioned to represent thing perceptually at all, and to assign this state a specific role (representing chairs) within that system. Because the perceptual system is reliable in the conditions which defined it as a perceptual system, its representative states are constitutively connected to truth, and so beliefs directly based on them, such as 'There is a chair before me' inherit not only the content of the perceptual representation, but their connection to truth, and agents are thus entitled to them. Even when the BIV says 'There is a chair before me', the belief it expresses is based directly on a perceptual state with a certain nature, and this nature ties it to a normal condition in which it reliably represents a physical environment. Thus the BIV believes that there is a chair before it, and is warranted in so believing.

[105] Burge 2003.
[106] Ibid, p 533.
[107] Ibid.

There is considerable agreement, therefore, between Burge and I about BIVs: they enjoy some positive epistemic status (perceptual entitlement, conformity with secondary norms of belief); they are in an abnormal case, contrasted with a normal case in which they interact with their physical environment; the normal case is what determines the BIV's content. There are also several differences between us. Burge focuses on the less extreme case of recent envattment, so it is not clear exactly what he would say about envattment from birth.[108] I use a different account of normality, which is general and glossed in terms of explanation. Burge pays special attention to *perceptual* representation, with the content of (perceptual) belief treated as falling straightforwardly out of an account of the former: when the transition between the two 'goes well', then 'reference is preserved'.[109] I treat propositional (and sub-propositional) representation directly, without reference to any other levels or kinds of representation.[110] But the crucial difference is that I have a general theory of content, whereas Burge does not. This is not to slight Burge: not every valuable enquiry into the nature of content need involve offering a general theory of content. The point is more that we have different goals, and our accounts do not clearly and directly conflict.

I believe, moreover, that my general theory predicts the more specific claims that Burge makes: or at least, very similar ones. First of all, the conditions that are normal on Burge's are also normal on mine, and vice versa. According to my theory, the conditions which fall within the explanatory order of human cognition (and so are normal, as I understand normality) are those which contribute to fixing the content of our terms and concepts (and so are normal, as Burge understands normality): content is determined by how much an agent knows across such conditions. Nor is this a superficial coincidence: Burge is very much

[108] 'If one is transported to an abnormal twin or vat environment, with no warrant for suspecting this…' Ibid p 538.

[109] Ibid p 541

[110] Also among the key differences, at least at the level of framing, between my view and that of Williams 2020.

preoccupied with the nature of representational states and systems, and providing biological explanations for the relationship between mind and world.[111]

Pushing on to specifics, recall that Burge has a recently envatted BIV base a belief on a perceptual representation as of a chair being before it, and this belief inheriting reference to chairs, as well as warrant, from the perceptual representation. As discussed, this is not quite how I explain the matter. But basing beliefs on perceptual experience[112] in this way evinces good epistemic dispositions: it leads to knowledge across normal cases, those falling within the explanatory order of human cognition, and so with reference to which the natures of our cognitive systems are defined. At least, it does so on an interpretation according to which the brain's chair-concept refers to the sort of thing that causes similar experiences in those normal cases. For that very reason, such an interpretation is correct, and the brain believes that there is a chair before it, and this belief conforms to an important epistemic norm. This upshot of my theory is, approximately, what Burge is saying. But note also that my theory is able to say something about other cases, such as those in which the transition between perceptual representation and belief fails to 'go well': content is not preserved because such beliefs do not conform to good epistemic dispositions. Now is not the time to attempt a thorough analysis of such cases and how they arise, but I believe that Madagascar-style examples[113] as well as more complex variants of the fake barn example above sometimes follow such a pattern. Cases based on singular reference will be discussed below. Albeit our explanations run in different directions, something like Burge's account of perceptual belief follows from my own theory.

Overall, my account of reference can provide a plausible error-theory for BIVs. Because they are disposed to acquire knowledge in normal cases by means

[111] Compare Burge's comments on developing a 'broadly Aristotelian framework' (p 509) with the Aristotelian epigraph of Williamson 2000: 'Everyone by nature desires to know'.

[112] I don't think it matters a great deal here, but we might prefer to think of this as quasi-perceptual experience.

[113] See Evans 1973.

of, for instance, their swan-concepts, envatted brains can refer to swans, and hold false beliefs about swans. Their capacity to refer to specific individuals is more limited, as expected, because it is hard to 'train' a concept on a specific individual without any perceptual links. While may theory does predict that some reference to specific individuals is possible, I think I have laid out plausible conditions for this to happen. This account can also subsume arguments about BIVs made by Tyler Burge.

B: The Swampman

I: Introduction

The Swampman seems much like us.[114] In some respects, he may be better than us, in that he is perfectly healthy: at least, he is a perfect replica of a healthy human being. But he is unlike us, in that he rose fully-formed from a swamp. Where I have an eye, he has a replica eye, shaped not by the exigencies of mutation and selection, but by sudden freak. Nonetheless, when there is a particularly treacherous bog before him, the Swampman's replica eye receives the light reflected from it and transmits signals along a replica optic nerve, and the Swampman's path adjusts to move safely around the bog. Does the Swampman believe that there is a bog before him?

This question matters because, on some views of content, content is defined by function, and on most views of function, function is defined by history. Biological function in particular is defined by evolutionary history.[115] The function of an ordinary human eye is sight; eyes have this function because in

[114] The swampman originates in Davidson 1987.
[115] Millikan 1989.

enabling sight they contributed to the fitness of hominids in the past, and that is the reason humans have them in the present. Using retinal scanning, an eye can also function as a key, but this is a function imposed from without, in the way that agents happen to use the eye. We are interested in *proper functions*, functions that are inherent to the eye, and which it would retain even if it were not used, or were broken and could not be used, in that way. The point is that these proper functions are dependent on history. Swampman's replica eye lacks a history; it simply popped out of the mire. Thus its proper function is not sight, or anything else. Function is foreign to Swampman.

Nothing in Swampman, therefore, has the function of representing the world around him, and nothing has the specific function of representing bogs. According to teleosemantic theories of content, this means that Swampman lacks content, and does not believe that there is a bog before him.[116] Given the similarities between Swampman and ourselves this may be a surprising claim: it looks a lot like he believes that there is a bog before him. Thus this case is often adduced as evidence against teleosemantics. Its advocates, however, accept the counter-intuitive result as the inevitable consequence of the best available theories of content and function. Swampman is thus an important test-case for theorising about content.

II: Normality for Swampmen

Since mine is not a teleological account of content, I will not be returning so direct a negative result for the Swampman. My own view is that the correct interpretation of a subject's cognitive system is that on which the agent is best disposed to know by means of that system. A simple response to the Swampman might be that he has a neural architecture much like ours, which underwrites a

[116] See eg Millikan 1996 for a discussion of how teleosemantic theories of content can handle swampman cases.

cognitive architecture like ours. Within that architecture is a concept which is activated when the Swampman's replica eyes receive light reflected from bogs, when bogs confound his footing, and so on. The agent will be best disposed to know by means of this concept if it refers to bogs, and so that is indeed its reference. The Swampman does believe that there is a bog before him.

Beneath the surface here, however, lies a slough of doubt. This response assumes that the Swampman has a cognitive system and is in a position to know about bogs. Both assumptions will be examined later. First, I shall review what optimizing an agent's disposition to know involves. An interpretation of a subject's cognitive system optimizes their disposition to know if, on that interpretation, the subject knows the most across normal cases at which they use the same cognitive system in the same way. At this point a worry arises: the very existence of a Swampman is abnormal. Recall above (5.A.III) that we defined normal cases for human beings in terms of deep evolutionary explanations. Chair experiences and light from chairs occur together by default, because human eyes and brains evolved together in a certain way. Without an evolutionary history, there is no similar explanatory order to define what is normal for the Swampman. How then are we to evaluate its dispositions to know?

While it is true that a Swampman comes into existence abnormally, it does not follow that there are no distinction between normal and abnormal cases for a Swampman. Explanations of what the Swampman does are perforce shallower than explanations for what ordinary human beings do, but he does fit into some explanatory order nonetheless. There are causal regularities in how the Swampman operates. He has replica legs, a replica brain, and a replica nervous system. There may be a distant abnormality in their origins, but once the causal mechanisms are in place, their functioning together provides a perfectly satisfactory proximal explanation for various events, such as the Swampman traversing round a bog. Sending electrical signals along his replica nerves to his replica legs to ensure that each replica foot falls successively in front of the other to exert pressure on the stable ground around the bog is the normal way for the Swampman to traverse a bog. If, on the other hand, the Swampman were spontaneously elevated and then levitated over the bog, that would be abnormal, because the complex causal regularities exhibited by the Swampman would have been suspended, and the explanatory order they define overturned. The Swampman's simply walking round is the default, the Swampman's levitating requires special explanation.

So too for epistemic cases. The replica eye and optic nerve and vision centre of the brain are all there, however abnormally they came about, and function together along specific causal pathways. The regularity within those causal pathways explain the Swampman's visual experience. If that experience is as of a bog, then light bouncing off the bog and onto the replica retina is the default case. Anything else requires a special explanation: maybe the bog-seemings are the result of mirage. In general, there was one distant, abnormal event: the materialization of the Swampman. This event established many mechanisms in the Swampman's body, which operate according to causal regularities just as the mechanisms within an ordinary human do. These regular mechanisms provide default proximate explanations for the various ways things appear to the Swampman. However impoverished these explanations are compared to the explanations available for ordinary humans, some pairs of appearance and reality fall into the Swampman's explanatory order, and others do not. These are epistemically normal cases, departures from which are epistemically abnormal. The correct interpretation of the Swampman is determined by how much he knows across such cases.

III: Cognitive Systems for Swampmen

We should now return to the first of the two assumptions noted earlier, namely, that the Swampman has a cognitive system. Anyone sympathetic to teleosemantics is likely to reject this assumption. The worry is that a cognitive system is a system whose *function* it is to represent the world, and the Swampman could have no such system: nothing about the Swampman is as it is because it contributed to the fitness of Swampman's ancestors. No history, no function; no function to represent, no representation.

I want to leave this option open for the teleosemanticist. My theory is foremost a theory of the interpretation of cognitive systems, and not a theory of the nature of cognitive systems. If the best theory of their nature is both teleological and etiological, then the Swampman is indeed without content. However, this is not a response that I favour. In particular, returning to a distinction drawn in 5.B.I, I would suggest that it suffices for something to *function as* a cognitive system, even if that is not its *proper function.* It is the proper function of a power cable to conduct electricity, even when it is broken; that function is inherent to the thing. Nonetheless, a broken power cable can function as a belt, even if that is not its proper function, because specific circumstances transpired that way: in this case, the cable happens to be supporting a garment around a person's waist. I would suggest that any system which takes in information and outputs aptly ordered (for organisms, let us say fitness-increasing) behaviour in a complex, open-ended way is a cognitive system. A system which takes in only nociceptor excitation and outputs only cries is not sufficiently complex or open-ended. A system which takes in either one of a range of visual patterns or one of a range of sonic patterns, and outputs either staying in doors or picking up an umbrella before going outdoors, is complex and open-ended. Ordinary human beings possess such systems, and so too does the Swampman. In this instance, the

Swampman's system took in light from the bog and put out a circuitous movement avoiding the bog.

One reason to suspect that mere functioning as cannot replace proper function is the possibility of malfunction. Human cognitive systems often go awry. We are often in the belief state that P even though P does not obtain, and there may be no behaviours clearly recognisable as having been guided by the assumption that P (for instance, the belief that Toronto is the capital of Canada). When a malfunctioning cognitive system lacks an evolutionary history, as the Swampman's does, what grounds its representations?

This may be a concern for molecular theories of representation, but mine is a holistic theory. I am primarily concerned not with individual concepts or belief states, but with whole cognitive systems. What may be the outright malfunctioning of a belief state is only imperfect functioning in the whole system. As discussed, a distinctive feature of cognitive systems is that they are complex and open-ended, and thus tolerant of local inaccuracy. Even if I believe that Toronto is the capital of Canada, my cognitive system may be functioning well overall.

In the case of a total breakdown of the cognitive system, when no new information is taken in, and no behaviours that could contribute to the organism's fitness are put out, then we might have a true absence of content. But crucially, my view would likely attribute content failure in such a case whether or not the proper function of a cognitive system is underwritten by etiology. So severe a failure of the cognitive system would probably be accompanied by a failure of knowledge. If the Swampman is simply not receiving light from the bog, or that light does not guide the Swampman's behaviour in ways that intelligibly serve his interests, then we have reason to deny that the Swampman believes that there is a bog before it, or is even in a position to know that there is a bog before it, independent of any considerations about the nature of cognitive systems.

Thus I am happy to apply a loose standard for the possession of a cognitive system to the Swampman. In fact, the Swampman has a system which takes in information and puts out behaviours that increase his fitness, and so he has a cognitive system which is amenable to interpretation. There is some component of this system which encodes the most knowledge across normal cases if it refers to bogs, and there is some more complex element, of which this bog-concept is a part, which is thus best interpreted as the belief that he has a bog before him.

The second assumption is that the Swampman is in a position to know. Independent of any considerations about cognitive systems, some argue that knowledge requires proper function on distinctively epistemological grounds. Alvin Plantinga has prominently defended such a view.[117] According to him, knowledge is true belief produced by the proper functioning of a subject's cognitive powers in an appropriate environment. Since we have supposed that the Swampman has no cognitive powers with defined cognitive function, he is not in a position to know anything. All potential interpretations of his cognitive system, therefore, leave him without the ability to know. Thus we should attribute total reference failure to the Swampman.

Proper function epistemologists are welcome to adapt this theory as they see fit, but such an account of knowledge is contrary to its spirit. This is a knowledge-first theory: it takes knowledge as explanatorily basic, rather than something else such as proper function. I take it that the best way of characterising the conditions in which knowledge obtains is *epistemic safety*: we tend to know that P if we cannot easily err that P. Granting that the Swampman is capable of holding beliefs, he is capable of holding safe beliefs: indeed, he is capable of producing safe beliefs by methods which tend to produce safe beliefs. The overwhelming presumption is that such beliefs constitute knowledge,

[117] Plantinga 1993.

whether or not they are produced by properly functioning cognitive powers. Thus the Swampman is in a position to know that there is a bog before him, and he will know the most in normal cases on the interpretation that he believes that there is a bog before him.

The natural response to the Swampman is that, if it sees more or less like a human, and it walks like a human, then it thinks like a human too. My account of content can vindicate this response. There are causal mechanisms at work within the Swampman just like those at work within an ordinary human, and these provide an explanatory order into which the Swampman's experiences can fall, defining epistemically normal cases for him just like the epistemically normal cases for an ordinary human. Somewhat tentatively, I suggest that the Swampman possesses a cognitive system even in the absence of proper functions. More confidently, I contend that the Swampman knows in the absence of proper functions (assuming that he does possess a cognitive system). There is a concept within his cognitive system by means of which the Swampman is best disposed to know if that concept refers to bogs, and so it does refer to bogs. The Swampman believes that there is a bog before him.

C: The Vatbrain

I: Introduction

Consider now the Vatbrain. Like the Swampman, the Vatbrain has popped instantaneously into existence. It is a perfect replica of a human brain, as the Swampman is a perfect replica of a whole human being. Unlike the Swampman, it is obviously unable to sustain itself unaided for any length of time. By a quirk of fate, however, it happens to have materialized within the laboratory of a brain-envatting scientist, right inside of a vat that is fully primed to receive, and deceive, a brain. Thus, not only is its continued existence secure, but it undergoes a simulation of ordinary human experience. We may suppose, indeed, that it undergoes the simulation of emerging from a swamp. Perhaps it is fed sensations as of being confronted by a particularly deep bog. Within it then occurs a duplicate of the neural activity by which an ordinary human being would redirect its steps around the bog, and the simulation device furnishes it with an appropriate experience. Does the brain believe that there is a bog before it?

The point of this example, of course, is to combine the distinctive features of two cases already discussed. Like the Swampman, the Vatbrain lacks an evolutionary history, and whether it has any biological functions is in doubt. Like the ordinarily envatted human brain, the Vatbrain lacks epistemic access to its physical environment. Thus contending with the worst of both cases, what content remains for the Vatbrain?

II: Cognitive Systems for Vatbrains

We should consider first whether the Vatbrain has a cognitive system. We assumed that the ordinary envatted brain does so: to make the point explicit, one of that brain's biological functions is to be a cognitive system. Meanwhile, let us grant the (admittedly tentative) earlier conclusion that the Swampman's quasi-brain functions as a cognitive system. Does the Vatbrain function as a cognitive system too? Recall that cognitive systems take in information and output appropriate behaviours in complex, open-ended ways, per 5.B.III. Now the Vatbrain does take in information. The simulation device directly feeds it information which is exactly equivalent to the sort of information that an ordinary human brain in an ordinary human body would receive, however oddly sourced.

Behaviour is more delicate. The Vatbrain cannot scratch its back or walk around a bog. As we've seen, though, it can undergo a duplicate of the neural processes which initiate such behaviours in human beings. Because of the way the simulation device works, these behaviours are arguably appropriate. Sending the back-scratch signals relieves the simulated itch. Sending the bog-avoidance signals will mean the simulation does not become imminently uncomfortable. This appropriateness may be conducive to fitness in a fairly strict sense. Suppose the simulation device runs on the Matrix protocol: death in the simulation means death in reality. Then vestigial bog-avoidance reduces the chance of indirect death by simulated drowning. Vestigial traffic-checking will reduce the risk of death greatly.

However, we might think that all this is too dependent on the caprice of the scientist. The vestigial behaviours are appropriate only because of the way the simulation device has been programmed to respond to them. Should the programming change, they will cease to be so. This is plausibly too weak a causal relationship to underpin a cognitive system.

Still, there is more to be said here. Many human behaviours are mostly mental. We can factorize numbers or compose poetry *in our heads*. These behaviours keep our minds sharp and, insofar as they are satisfying, relieve stress. Thus they increase our fitness. Admittedly such behaviours are very loosely related to the intake of new information: one might factorize numbers on a mere impulse. Consider a more complex case. Sally suffers from anxiety. She is apt to fall into negative patterns of thought. She tends to think that people dislike her, and will respond with hostility when she joins a social group. She also thinks that the university buildings are sterile and overbearing, and hates being in and around them. Sally has heard that challenging such thought processes might be good for her mental health. In fact, she has noticed several previous occasions when she was consumed with dread before a social event at which she expected a hostile reception, only to be made welcome when she arrived. How much better off would she be had she challenged her negative thoughts from the start? In light of such information she adapts the new behaviour of challenging her negative thoughts. She tells herself that people like her, and will interact with her in a friendly manner when they meet. She tells herself that the university buildings are not so much overbearing and sterile as spacious and clean, and attends to different features of the buildings which better suit this description. As a result, she starts to feel less anxious: her mental health has improved and her fitness has increased.

Vatbrain could struggle with similar anxiety and respond to it in a similar way. All the information is still being fed to it by the simulation device, but it is, we may suppose, making a rational inference from introspection, induction, and serviceable psychological heuristics, and thus adopting a behaviour that benefits mental health in an ordinary and direct way. It is training itself to think in ways that tend to benefit human brains: and thus, given the resemblance, itself. The brain is still in the scientist's power, and so he could interfere with the causal process at any time, but it is an independent causal process which he may choose to interfere with, rather than a deliberately engineered one that he is responsible for arranging. The relationship between the behaviour and its positive outcome depends on the scientist’s whim no more than any of the steps that we ordinarily take to help ourselves depends on our not being randomly murdered. There is a *prima facie* case that the Vatbrain does possess a cognitive system.

III: Interpreting Vatbrains

Granting this, the central question is how to interpret the Vatbrain's cognitive system. Depending on the assumptions used to fill out my basic framework, there are three options: error-theory, phenomenalism, and reference failure. My view is that the correct interpretation of a subject's cognitive system is that on which the subject is best disposed to know. How well disposed a subject is to know is determined by how much they know across normal cases. The question, once again, is what is normal: in this case, normal for a Vatbrain. What sort of explanatory order applies to it? For a standard BIV, the thought was that it is a human being, and so whatever is normal for a human being is normal for an envatted brain. A Swampman is not quite a human being, but that thought was that it contained complex causal mechanisms, and the continuing functioning of those mechanisms is normal, and any interruption of them would be abnormal. Because those causal mechanisms replicate human ones, the end result is that what is normal for a human being is also normal for a Swampman. The Vatbrain has some of the causal mechanisms, of course, but not all of them. What makes it the case that there is normally a bog before the swampman when he seems to see a bog is the cooperation of the replica eye and replica optic nerve and replica brain. Since the Vatbrain is a replica brain and nothing more, without any replica organs able to transmit visual information to it, there is not normally a bog before it when it seems to see a bog.

An error theory along the lines of that we endorsed for the standard BIV thus seems unpromising. Given the *prima facie* case we have already considered for the claim that the Vatbrain does possess a cognitive system, however, we should make some attempt to interpret it before defaulting to the option of reference failure. While the Vatbrain lacks the three-way interaction between the Swampman's brain, body, and environment, it has its own equivalent of this: the interaction between the brain, the life-support systems sustaining it within the vat, and the simulation device feeding it experiences. The continuing operation of these causal mechanisms defines the explanatory order into which the Vatbrain falls, and, in turn, the normal cases for the Vatbrain.

Is the Vatbrain in a position to know anything in these cases? I believe so. The simulation device has been programmed to ensure a coherent stream of experiences for the Vatbrain. Accordingly, it keeps track of Vatbrain's apparent chair-sightings, so that it will continue to provide chair-experiences at appropriate points. Likewise, my computer keeps track of my changes to the Vatbrain discussion, so that it will continue to display the correct images at the appropriate points. When this system of storage and retrieval is in place, we say that my paper exists as a file on my computer. My computer may even keep track of when to display red chair-images, if I am playing a game which features red chairs. We say that there are (virtual) red chairs in the game.[118] Accordingly, we might interpret the Vatbrain as thinking about virtual chairs: computational entities of the same familiar kind as word processing documents and video game chairs.[119]

This approach does have its own limitations. The simulation-device presumably does not run a comprehensive simulation of the entire universe. It provides what David Chalmers calls an 'extendible local matrix'. Suppose the Vatbrain says to itself 'There are more red chairs in the world than I will ever see'. Not all the red chairs that the Vatbrain will see red chairs have yet been

[118] Should virtual red chairs be assimilated to fictional red chairs? I think not. There may be a thousand red chairs in the castle of the Red King according to the game's fiction (its 'lore'), while only the hundred in the main hall are accessible to the player. So there are fewer virtual chairs than fictional chairs. Moreover, games need not involve fictions: there is no clear sense in which tetriminos exist according to the fiction of tetris, or that pawns exist according to the fiction of an online chess game.

[119] Chalmers 2005 defends such an approach for ordinary envatted brains.

stored in the device’s memory, and few, if any, other red chairs are so stored. What, then, is the Vatbrain counting?

There are two main alternatives here: perhaps there are no chairs which it will never see, and so it is speaking falsely. We will thus only be able to interpret the Vatbrain as partially knowledgeable, falling back on an extensive error theory for anything beyond what has been specifically programmed into the simulation. Otherwise, we might try to invoke counterfactuals about what the simulation device would do. There are more virtual chairs that the simulation device would show the Vatbrain than those that it will show. The difficulty here is that the spectre of merely possible virtual chairs now looms: chairs that would exist within the simulation had the virtual carpenter lived longer.

At this point we are working our way back to some kind of error-theory, which is promising, because, much like a standard BIV, it looks like the Vatbrain is radically mistaken. The precise extent of its knowledge is a fine question we need not settle here. What matters for the overall discussion is that we have a strategy for interpreting the Vatbrain: it refers to the computational objects it knows about in normal cases. It does have some content, however limited. My theory of content thus accommodates three intuitions about Vatbrain. It seems to have some beliefs, and so indeed it does. Lacking both a full human body and a biological history, it is very different from us, and thus we might expect its content to be different too, which it is. Finally, the Vatbrain seems mistaken about what the world is like and, given the limited resources available for interpreting it, its beliefs are mistaken

The Vatbrain probably possesses a cognitive system. In order to interpret it, on my view, we need to determine which cases are normal for it, and what it can know in those cases. The normal cases for it are those in which the causal mechanisms within it, and with which it interacts, continue smoothly. In those cases, it knows about computational objects of the simulation. This is the basis for the Vatbrain's content. Because the simulation is only an extendible local matrix, while the Vatbrain replicates the neural structures of a human being with beliefs about a large permanent physical universe, many beliefs that the Vatbrain has are likely to be false.

The optimizing dispositions to know account of content is well-equipped to deal with intriguing edge cases, ones with which alternatives may struggle. For BIVs, it is able to provide an error-theory, avoiding both anti-realism and complete reference failure. Such a verdict is not easy to reach on other variants of Interpretation, including simple knowledge-based theories. For Swampman, it is able to vindicate the standard intuition that the Swampman has content just like an ordinary human. Although I consider this a strength of the theory, it still depends on certain assumptions, which could be rejected in order to open a path to reconciliation with teleosemantic approaches. Finally, it offers the beginnings of a plausible response to the hybrid Vatbrain case. Since this scenario is so extreme, we ought not to set too much store by it, but it is to the theory's credit that it has something sensible to say in such a pinch. The optimizing dispositions to know theory holds its own at the edges of interpretation.

Chapter 6: Interpretation and Practical Reason

One potentially surprising feature of knowledge-based theories of content is that they take only inputs, and not outputs, into consideration. That is, the interpretation of an agent revolves entirely around their beliefs and not their actions. Many Principles of Charity stress both. Rationality-based accounts, in particular, often try to maximize the practical as well as the theoretical rationality of agents. Accordingly, I will investigate whether anything is lost when removing action from the basis of interpretation. First I will consider general reasons for basing interpretation on action, and then I will consider the case where a critical role for action looks most likely: moral concepts.

A: Action and Interpretation

I: Desire as Determinant

Suppose you are an anthropologist doing fieldwork. You have learned a little of the local fauna. There are two species of snake: the more numerous species is harmless, but the rarer one is poisonous. They look very similar, but can be distinguished at a distance by subtle markings, if one is paying sufficiently close attention. You are now taking a short trip with a member of the tribe you have just contacted, and whose language you are trying to interpret. You have not yet had any opportunities to study their snake vocabulary, but a little way ahead one crosses your path. Your companion glances at it quickly, and utters ‘Yorum’. What precisely is the meaning of ‘Yorum’?

Well, presumably your companion knows that a snake crossed his path. Let us thus assume that he was referring to the snake, although this leaves much unsettled. Does ‘Yorum’ refer to snakes in general, to the harmless species, or to the poisonous species? You know that he was, broadly speaking, in a position to know which species of snake it was, though you can’t be sure whether he made the most of that position. In an even broader sense, you yourself are in a position to know which species the snake is, inasmuch as you can run up and get a clearer view. But this will not tell you anything about the specificity of your companion’s knowledge, nor the specificity with which he chose to speak. Reflections on his epistemic position regarding the snake will only go so far in helping you interpret his use of ‘Yorum’.

There are, however, other kinds of evidence that we have not yet considered. Suppose the utterance of ‘Yorum’ was loud, and apparently distressed. Suppose that your companion became perfectly still, and indicated for you to do likewise. Suppose, in short, that we look not only at your companion’s ‘inputs’ - how long he was looking at the snake, how good a view he had, etc - but also his ‘outputs’ - the sort of actions he took on the basis of the information he had just received. These outputs may be decisive where the inputs are ambiguous: just the sort of action you would expect if ‘Yorum’ referred specifically to the poisonous species of snake. That, one might suppose, is thus the correct interpretation of ‘Yorum’. One might further suppose that the correct method of interpreting your companion involves reference to his outputs.

This would appear to put rationality maximizing approaches to interpretation at an advantage. For the rationality of belief is only one aspect of rationality: actions can also be rational. Some theories of interpretation aim to maximize rationality of both action and belief. Whether ‘Yorum’ refers to snakes or poisonous snakes, your companion may well have expressed knowledge. But the action of stopping and standing still is more rational given that he believed specifically that there was a poisonous snake. Rationality maximization, unlike knowledge-based approaches, thus offers a decisive verdict in favour of the poisonous snake reading. It is therefore the superior approach to interpretation.

The knowledge theorist might rejoin that this is an artifact of overdramatisation. As a practical matter, we would indeed want to look at outputs in order to learn what is the correct interpretation of your companion. But the way we *learn* about correct interpretations is a secondary point. The main point is what makes those interpretations correct in the first place: per Lewis, how the facts determine the facts. The knowledge theorist maintains that it is facts about knowledge that determine facts about interpretations, while facts about action do not. But facts about action may still be *evidence* for facts about knowledge, and so may be practically relevant for our attempts at interpretation.

Considering only the circumstances in which he formed his belief, that is, his inputs, your evidence for your companion's knowledge was ambiguous. Considering his outputs yielded more evidence *about his knowledge*. He cried out and stood still because he knew the snake was poisonous. That is why we should think that 'Yorum' refers to poisonous snakes. More generally, the knowledge theorist maintains, knowledge is the norm of action. So knowledge and action are closely connected: the search for a belief that rationalizes a given action is *ipso facto* a search for an item of knowledge, if the action is to be fully rationalized, or at the very least a belief that evinces a disposition to know, if the action is to be excused or weakly rationalized. Thus attending to the actions of any subject will yield important evidence about their knowledge and dispositions to know. The knowledge theorist need concede no ground to the rationality theorist.

II: Rationality of Desire vs Rationality of Belief

Might desire constitute content, in addition to aiding acts of interpretation? According to the rationality theorist, the correct interpretation of an agent maximizes the rationality of both their beliefs and desires. If this is right, we should be able to construct cases in which slightly less rational beliefs are traded off against drastically less rational desires, where the correct interpretation preserves the rationality of the agent's desires. Consider the following case.

Case: Friendly Fire

> Miles is a soldier in the midst of battle. Most of his enemies wear yellow uniforms, as do a small number of his allies. His unit is under heavy fire, and receiving reports of enemy advances across the field. A soldier in a yellow uniform approaches Miles. Miles shoots. The soldier turns out to be an ally.

Before moving on to our analysis of this case, we should take the time to unpack a few points. First, I am using a simplified model of motivation on which actions are fully explained by the interactions between belief and desire. This makes the problem vivid and tractable: I trust that what I say with respect to this simple model can be generalized to cover more complex alternatives, until reasons are given to think otherwise. This means I may talk blithely about Miles desiring to shoot an enemy, when we might imagine that there is an important sense in which Miles desires no such thing: perhaps Miles will be consumed by regret for the rest of his days. The belief/desire model has its ways of dealing with this: perhaps there is a conflict between Miles's more basic desires (say, the desire to do no harm, and the desire to protect himself and his comrades), and this internal battle has been decided, at a psychological cost, in favour of shooting enemies, a situation we describe through the shorthand ‘Miles desires to shoot an enemy’. As always, we are simplifying so that the salient theoretical points are easier to explore.

A further issue is that we will be directly evaluating the rationality of desires. There is some tension between this point and the last, since a view known as neo-Humeanism holds both that agents are exclusively motivated by desires *and* that actions are exclusively rationalized by desires.[120] Only desires give agents reasons for action, and desires themselves are not themselves rational or irrational: in Hume’s famous phrase, ‘reason is, and ought only to be, the slave of the passions’.[121] Indeed, if this Humean view is true, then the rationality of desire cannot influence interpretation. For the sake of the argument, however, I am addressing the view found in Williams, where actions are explained by the interaction of belief and desire, and desires can be evaluated for their substantive rationality.

Returning to the case in hand. Did Miles believe his target was an enemy, and desire to shoot an enemy, or did he believe the target was an ally, and desire to shoot an ally? It may be more rational to believe that an ally is an ally, rather than an enemy, but it is understandable how such a mistake might have been made. Miles was under pressure, and soldiers in yellow uniforms are more often allies than enemies. A false belief, in this case, is only a little irrational. But the desire to shoot an ally is simply perverse. We maximize Miles' overall rationality, therefore, by attributing him the belief that his target is an enemy, paired by the desire to shoot an enemy. Moreover, what makes this interpretation correct is the rationality of the desire.

[120] See Sinhababu 2017 for a recent defence.
[121] Hume 2000 [1740], III.III.III

I take it that rationality maximization yields the right verdict in this case: the unfortunate Miles really was acting on the false belief that his target was an enemy. The question is whether this verdict need be reached via the rationality of desire. A simple truth-maximizing view would struggle here: given a binary choice between the true belief that the target is an ally, and the false belief that he is an enemy, it is bound to favour the former. But knowledge works differently. For the belief that the target is an ally, while true, is not plausibly knowledge. The story turns on Miles being easily mistaken. Hence his belief, even if true, would not be safe, and so it would not be knowledge. Since Miles is in a poor position to discriminate between yellow-uniformed allies and yellow-uniformed enemies, attributing him the belief that the soldier is an ally would not advance either his knowledge or his dispositions to know, and so a knowledge theory would not favour the mistaken truth-maximizing interpretation.

To apply more pressure, we might consider some more complex alternative interpretations. What if we construe Miles as believing that his target is wearing a yellow uniform, and desiring to shoot someone wearing a yellow uniform? Given the description of the case, we might assume that Miles is in a position to know that his target is wearing a yellow uniform. So knowledge maximization should favour this new interpretation. Nonetheless, it still looks like the enemy interpretation is more plausible. What makes that interpretation plausible, again, is that it attributes a more rational desire: the desire to shoot an *enemy*, rather than, capriciously, just someone who happens to wear a yellow uniform. This gain in the rationality of desire trumps the gain in knowledge, *contra* the knowledge theorist.

Once more, however, this is not the only diagnosis available. We might first note that the two different belief-attributions are not, of themselves, exclusive. The soldier might believe both that his target is an enemy and that he is wearing a yellow uniform. Tidying the case up a bit, then, let us suppose that we have isolated the belief-state that bears direct causal responsibility for the action. So the question becomes what the content of *that* belief-state is. Even so, having first acknowledged the compatibility of both beliefs in principle relieves the pressure on knowledge maximization considerably. By identifying **My target is an enemy** as the content of this belief-state, our overall interpretation of the soldier need not lose him the knowledge that his target wears a yellow uniform.

Some loss is unavoidable of course: the specific belief-state which interests us encodes ignorance rather than knowledge. That is fine, however: knowledge theories are consistent with the plain fact that our beliefs sometimes fail to constitute knowledge. Attributions of ignorant belief are underwritten by the principles of compositionality. Take the concepts which compose the belief-state. Among them will be one whose content we interpret as either A **is an enemy** or B **is wearing a yellow uniform**. Considering the occurrence of this concept throughout Miles's entire cognitive architecture, the enemy interpretation will yield far more knowledge, and far better dispositions, than the yellow-uniform interpretation. He believes that *those ones* are a threat to him and his comrades; that *those ones* are advancing across the field; that Napoleon was one of *those ones* to Wellington. These beliefs are knowledge if *those ones* are enemies, and not if *those ones* are wearers of yellow uniforms. Knowledge theories, therefore, favour the correct interpretation after all.

To see more clearly the principle behind this verdict, it may help to consider variant cases. In one case, Miles is able to acquire overwhelming evidence that the yellow-uniformed target is an ally. After double-checking his evidence, he proceeds to shoot anyway. In another, Miles was given specific orders about how to respond to risk in the uncertainty of the battlefield. Yellow uniforms are so likely to be hostile, his commander said, that the best response overall, given the stakes and the difficulty of acquiring and assessing information, is to shoot on sight. Once battle is joined, he shoots a yellow-uniformed ally.

In the first case, I take it that Miles believes that his target is an ally. Betrayals sometimes happen: given minimal information, the natural assumption is that it would be irrational to turn against an ally, but of course there are ways to fill out the case so that such an action would be weakly (bribes?) or even strongly (conscription into an unjust cause?) rational. After a certain point, even the most practically-minded rationality theorist will admit that the epistemic irrationality of continuing to believe the soldier is an enemy comes to trump the *prima facie* practical irrationality of shooting an ally.

In the second case, meanwhile, I take it that Miles believes that his target is wearing a yellow uniform, and that this belief plays a decisive causal role in his action. He may not *want* to shoot at yellow uniforms in any psychologically robust sense, but as we noted, he may not *want* to shoot at anyone at all. Nonetheless, as with desiring to shoot enemies in the initial case, it is intelligible how the orders he received have shaped his motivational structure such that we can reasonably attribute him a desire to shoot at those wearing yellow uniforms, at least in the thin theoretical sense in which a desire is whatever combines with a belief to produce an action.

A knowledge theory can easily diagnose the differences between these cases. In the first case, the epistemic situation is different: Miles is in a position to know that his target is an ally. Therefore, attributing that belief maximizes his knowledge. In the second case, there is a difference in Miles' cognitive architecture. Hearing his orders activated a concept which encodes the most knowledge if it refers to yellow uniforms. In the subsequent battle, that concept will be salient, entering into new beliefs which may play a key causal role in acts of shooting.

Considering ourselves as potential interpreters, it is true that we would never perform such complex operations. Our understanding of the connection between belief, action, and desire provides a shortcut. Given our background psychological knowledge, and the way the case is described, we can deduce that Miles' cognitive architecture must have a certain sort of profile: this goes hand in hand with our pre-theoretic judgements about each case. In the original case, our judgment is that he believes that his target is an enemy; naturally, we expect him to believe that enemies are a threat to him and his comrades; insofar as we presume thought to be compositional, we expect there to be recurring components across these and a range of other beliefs. What makes our preferred interpretation correct may thus still be that it maximizes knowledge across Miles's cognitive system, even if we cannot do the relevant cross-referencing ourselves. Again, desire can play an epistemic role in assisting our acts of interpretation without playing a constitutive role in determining content.

B: Knowledge and Stability for Moral Concepts

We have seen that there is no general reason to assume that action need form part of the basis of interpretation. There is still a specific class of cases that needs to be addressed, since they are of such great overall importance, and seem to be closely tied to practical reason. We need to show that a knowledge-based theory of content can handle moral concepts adequately. Moral terms and concepts seem at least somewhat stable. Despite deep differences in beliefs and practices, surely most human communities are engaged in recognisably the same enterprise of moral evaluation. Most modern Britons would assent to 'Infanticide is wrong'; an Ancient Spartan would dissent from an appropriate translation. Here, we think, is a disagreement about what is wrong. This is not like the Briton and the Spartan 'disagreeing' about what is happening next month because they use different calendars to divvy up the year: they have made genuinely conflicting moral evaluations. They dispute whether to ascribe the very same property of wrongness to the very same action. What makes it the case that moral concepts show this kind of stability, referring to the same properties despite so many underlying differences in their circumstances of use?

I: Rationality and Moral Twin Earth

Consider two communities, Kantsberg and Utilitopia.[122] Each speaks what is recognisably a clone of English, but there is no contact between the two. Everyone in Kantsberg is a convinced Kantian on moral matters. They consistently say ‘That’s good’ of actions that fulfill the categorical imperative, and ‘That’s wrong’ of actions that violate it. In Utilitopia, by contrast, everyone is just as convinced a consequentialist. They consistently say ‘That’s good’ of actions that cause the greatest happiness for the greatest number, and ‘That’s wrong’ of actions that reduce happiness. Given the clear differences in their usage of ‘good’, then, why should we suppose that the word means the same thing in the language of each? In particular, why doesn’t ‘good’ just mean *fulfilling the categorical imperative* in the mouths of Kantsbergers, and *increasing happiness* for Utilitopians?

The rationality-maximizer has a clear answer for this.[123] He can appeal directly to moral reasons as a component of his overall conception of rationality. Williams argues that attributing blame if and only if wrong has been done is more rational than attributing blame if and only if some other condition has been fulfilled: violating the categorical imperative, for instance, at least assuming for argument’s sake that utilitarianism is true. So if an agent is disposed to attribute blame to another if and only if they would apply the word ‘wrong’ to them, then their dispositions are more reasonable if ‘wrong’ refers to wrongness than if it refers to violating the categorical imperative. Therefore, according to rationality

[122] The ‘moral twin earth’ case was introduced in Horgan and Timmons 1991.
[123] I am here summarising the answer given in Williams 2020, Chapter 5.

maximization, the agent’s word ‘wrong’ refers to wrongness, other things being equal.

This is how things are in Kantsberg. The Kantsbergers, good Kantians that they are, use their word ‘wrong’ to distinguish actions that violate the categorical imperative from others. But this is just one facet of their use of the word. It is tied to a complex of beliefs and practices, including the attribution of blame, which mark out its conceptual role. Whenever they use ‘wrong’, they assign blame, and whenever they retract the term they retract the blame too. This is in fact central to the term’s usage, while the articulation of Kantian moral theory and its application to specific cases, though entrenched, is more peripheral.

If ‘wrong’ in Kantsberg English refers to wrongness, then Kantsbergers are less than fully rational to the extent that they consistently apply to it violations of the categorical imperative (recall again that we are assuming the truth of utilitarianism). But they’re not awfully irrational: Kantianism is a well-developed moral theory with considerable merit. Many thoughtful people with normal powers of moral reasoning endorse it on considered reflection, and more will at least appreciate its initial appeal.

If, however, ‘wrong’ refers to the violation of the categorical imperative, then the Kantsbergers will be less than fully rational for the reasons discussed above: it is not reasonable to blame someone for violating the categorical imperative in cases where doing so is not morally wrong. Insofar as Kantianism is somewhat rational, as we have seen, then it may be somewhat rational to blame someone for violating the categorical imperative without doing wrong: the blaming inherits the rationality of the Kantian theory. Crucially, however, this is not how the term is being used: the attribution of blame is central, the expression of Kantianism peripheral. There is thus no question of inheriting the rationality of Kantianism: the Kantsbergers would simply be wedded to blaming violators of the categorical imperative come what may, which is irrational. If ‘wrong’ refers to violation of the categorical imperative, there is a deep irrationality at the centre of its usage. If it refers to wrongness, there is a minor irrationality on the periphery. Thus rationality maximization predicts that ‘wrong’ refers to moral wrongness in Kantsberg English, despite the Kantsbergers’ mistaken Kantian consensus.

II: Knowledge and Moral Twin Earth

Knowledge maximization, however, appears to struggle here. In most of the cases where Kantsbergers utter sentences of the form ‘F-ing is wrong’, they are in a position to know that F-ing violates the categorical imperative, and in far fewer are they in a position to know that F-ing is wrong. As we have seen, there is more to their use of the term than this. Kantsbergers will also say ‘If he did wrong, then he should be blamed’ and so on. The trouble is that this is more Kantsberg English, demanding to be interpreted. Perhaps in their mouths ‘should’ has a special Kantian sense which makes ‘If he did wrong, then he should be blamed’ true in Kanstberg English. After all, Kantsbergers will say ‘You should F’ even when you shouldn’t F, if F-ing fulfills the categorical imperative. Given how the whole range of Kanstberg’s apparently moral vocabulary is used, it seems it would express the most knowledge if it bears special Kantian meanings.

I am not primarily interested in simple knowledge maximization, however, but rather in optimizing dispositions to know. This means I want to maximize knowledge across normal cases. In close cases, the Kantsbergers’ use of moral vocabulary has a strong Kantian inflection, but I contend that there are plenty of more remote normal cases where moral vocabulary is used in the same way, but the Kantian inflection is absent. In these cases, the vocabulary is still being used in the same way because its function in sincere, reflective moral enquiry is core to its usage. If this were not so, we would not be so quick to assume that their word ‘wrongness’ refers to wrongness in the first place. It just so happens that, among the Kantsbergers, this enquiry has issued in firm Kantian conclusions. Suppose, however, that we are optimistic about moral enquiry: it tends to produce knowledge. In which case (at least continuing on with our earlier assumption that Kantianism is false), Kanstberg must be seen as a rarity. There may not be anything abnormal about Kantsberg - actual human communities come to moral agreements, and, like modern Britons and ancient Spartans, sometimes the communities conflict - but it is not representative of normality. In many more cases in which the specific conditions that created the Kantsberg consensus do not hold , the citizenry would abandon their Kantianism and come closer to the moral truth. Since moral evaluation is central to their usage, differences in the outcome of that evaluation notwithstanding, they would express that truth using the same vocabulary in the same way as they do in Kantsberg. Across normal worlds, therefore, Kantsberg English expresses more knowledge if ‘wrong’ refers to wrongness rather than mere violation of the categorical imperative.

The assumption of optimism about moral enquiry here is no defect. While some may not share such optimism, it is a natural partner to referential stability. Both are part of what we might think of as a high view of morality: there are special, mind-independent moral properties, which do in fact undergird our moral reasoning, discourse, and practice. Whether we should accept this view is one of, if not the, central issue in meta-ethics, and it would be surprising if we could settle it by working out a theory of content. It would be altogether less surprising, however, if it revealed new connections between the component of the high view under its direct aegis - referential stability - and other components of the view - in this case, optimism about moral enquiry.

III: What is a Core Disposition?

So far, I have still only sketched my response to the twin-earth problem. First, I will try to motivate the general claim, which might be suspected of question-begging, that the Kantsbergers' Kantian dispositions are not central to their usage of moral vocabulary, and so not decisive for interpretation. What right have I to assume that the Kantsbergers might abandon their Kantianism, and yet continue to use the same moral concepts in the same way? Given that my theory is about optimizing dispositions, it seems relevant that the Kantsbegers are strongly disposed to make Kantian judgements. In highlighting cases where the Kantsbergers stop making such judgements, this worry goes, I am not so much optimizing their dispositions as I am ignoring them.

Let us consider now a third community, that of Rosstov. Whether they would put it to themselves in quite this way, a moral philosopher observing their attitude to lying might say that they recognise a *prima facie* duty not to lie. Their default assumption is that a lie is wrong, but they accept that special circumstances in which it is not *might* apply. Once upon a time, Rosstov is visited by Professor Evil. 'Professor Evil' is not the name by which he is actually known, but rather a rough translation of that name into English from his native language. Rosstovians are thus not apprised of Professor Evil's evil. Having, as established, a general presumption against lying, they are disposed to make judgements according to the following pattern: if *A* is lying to Professor Evil, then *A* is doing wrong. But one day a brave hero uncovers the dastardly schemes of Professor Evil. The Rosstovians now believe that speaking truly to Professor Evil will aid those dastardly schemes, and lying to him will thwart them. Accordingly, their dispositions change, and they begin to make judgements according to another pattern: if *A* is lying to Professor Evil, then *A* is doing right.

Has there been a change in the meaning of Rosstovian moral terms, such that lies told to Professor Evil once fell into the extension of their predicate ‘is wrong’, but now fall into the extension of ‘is right’? Evidently not. This shallow change in dispositions does not make for a change in meaning. We turned to dispositions in the first place because we wanted to avoid ‘noise’ of exactly this sort from influencing interpretation. The important disposition here is the deeper general disposition to treat all lies as wrong *prima facie*. That disposition is the signal, which dominates the weaker disposition to treat lies to Professor Evil in particular as wrong. The latter is a mere artifact of how the particular circumstances of a case (ignorance of Professor Evil’s evil) affected the application of the deep disposition. To put the point more precisely, we might say that the one disposition is *explanatorily prior* to the other. When trying to optimize an agent’s disposition to know, what matters are the most *explanatorily basic* dispositions. The explanatorily less basic or posterior dispositions, such as the original disposition to judge lies told to Professor Evil as wrong, are noise which is properly ignored. Interpreting Rosstovians aright requires us to consider all the cases in which the basic disposition is followed, not just those where the less basic disposition is followed too.

Unable to proceed with his dastardly schemes in Rosstov, let us suppose that Professor Evil now travels to Kantsberg. Of course, the Kantsbergers are disposed to form judgements according to the pattern: if *A* is lying to Professor Evil, then *A* doing wrong. Once more, Professor Evil’s schemes are uncovered, and, as expected, Kantsbergers retain their earlier dispositions. Then a further change occurs: the Kantsbergers moderate their Kantianism to accept that lying is justified in certain conditions, including ones fulfilled in speaking to Professor Evil. Just like the Rosstovians, the Kantsbergers are disposed to form judgements according to a new patttern: if *A* is lying to Professor Evil, then *A* is doing right. The question arises: has the content of moral language changed among the Kantsbergers any more than among the Rosstovians?

One way to answer affirmatively is to stress that the Rosstovians changed when they learned more about *the facts*, changing their judgements only as they came to accept new descriptions of objective, empirical reality. The Kantsbergers on the other hand changed their judgements not according to the facts, but according to a change in values. This explains why meanings changed for the Kantsbergers but not for the Rosstovians.

In the first instance, we might reply that this diagnosis is misleading. The relevant fact that made Rosstovians change their mind about lying to Professor Evil was the fact that his scheme was *dastardly*. Perhaps we could substitute a more neutral description of what the bad Professor was up to: say, compelling orphans to moderate social media networks. But how does following the disposition to judge that *lying is right if the prima facie duty not to lie is outweighed* yield the more specific disposition to judge that *lying is right if it is done to hinder Professor Evil from compelling orphans to moderate social media networks?* Presumably through some general account of what sort of considerations outweigh the duty not to lie, which will itself be strongly value-laden.

Even supposing that we may neatly separate the factual and evaluative elements in the reasoning of the Rosstovians, though, and find conclusively that they only changed their dispositions to judge because they changed their factual beliefs, this response is still ultimately question-begging. According to what I have called the high view of morality, there are moral facts out there for the Kantians to discover, just as they had previously discovered the facts of Professor Evil's scheme. It is not merely a matter of carving up the same empirical facts according to a new convention, so that the change in judgement must issue from a change of meaning. Instead we might think that something about the world is learned by consulting moral sentiments, identifying presuppositions of practical reason, or practicing specific virtues. Whatever the details, the high view holds that there is a substantive process of moral enquiry which tends to yield knowledge. If that is so, there is no reason in principle why the Kantsbergers's moral language should have undergone any greater change in content than the Rosstovians'.

The less tendentious way to argue for a difference between the two cases is just to note that the Kantsbergers are overturning deeper, more entrenched dispositions. Speaking of Rosstovians, we had allowed that the disposition to treat all lies as wrong *prima facie* was the signal. But the equivalent of that for the Kantsbergers is the disposition to judge any and all lies as wrong. So surely the Kantsbergers have changed the signal outright.

Well, that will depend on how far back we think the real signal is. Suppose that they decide that lying is consistent with an overall Kantian framework, as Korsgaard, for instance, has argued.[124] Then the change of dispositions no longer seems so profound. Judgements about lying are downstream of judgements about whether actions follow the categorical imperative, which are explanatorily prior. The community has reconsidered whether lying is a violation, and accordingly altered their patterns of use, but the meaning of the moral terms remains just as consistent as with the Rosstovians.

IV: What is the Correct Interpretation?

Now it should seem at least somewhat plausible that Kantsbergers might mean the same thing by their moral vocabulary across cases where their specific judgments sharply differ. To make good on my overall point, however, I had better produce a serviceable account of what the 'real signal' is. What are the explanatorily basic dispositions behind the Kantsbergers' use of moral vocabulary? As suggested, I think the basic disposition is to use this vocabulary in connection with moral enquiry. But what does this mean?

[124] Korsgaard 1986.

A single party might be undecided about their course of action. Two parties might prefer two different courses of action. Such quandaries are resolved through some process of practical deliberation, deliberation about what to do. Turn left, or turn right? Typically, people will review reasons for going left, and reasons for going right, and try to weigh them against one another and make a decision, a decision to take one course of action instead of alternatives. Much of the time such deliberation takes place in a straightforward, piece-meal way. All parties agree on a goal, that of reaching an oasis quickly and safely, and the only question is whether that goal is more likely to be achieved by turning left or right. In more complicated cases, there might be agreement on goals, and perhaps agreement about efficacy of means, in some sense, but a reason of a different kind is allowed to settle the matter. Perhaps the oasis on the left is nearer, and there are fewer physical dangers en route, but the journey would take the party over an ancestral grave, and that is Taboo. So the decision is to turn right. Eventually, however, high-order practical questions tend to assert themselves. Is our goal really to be a safe journey to an oasis, rather than an arduous, but potentially, more rewarding, journey to a river? Is the old taboo over ancestral graves still to decide such questions on its own? So we are involved in *foundational practical deliberation*, where we deliberate about the very basis of practical deliberation, what primary goals to have and what reasons to prioritize.

Foundational practical deliberation is a stubborn fact of, at least, the human condition. Unless we reach what English speakers would call moral consensus, it is intractable. Some will continue to weigh reasons differently from others, or be unsure what goals to adopt: for this situation to be resolved is just for there to be a moral consensus. This foundational practical disagreement is what I call moral enquiry. We can identify agents and communities as engaged in moral enquiry without assuming that they use vocabulary or concepts even roughly equivalent to our own moral vocabulary and concepts, just by recognising them as deliberating over their choices of action in a sufficiently complex way.

Many different kinds of vocabulary could, in principle, be vehicles for such enquiry. Perhaps a group could use many descriptively rich ‘thick’ moral terms, similar to our ‘cowardly’ or ‘loyal’, without any ‘thin’ general terms such as ‘good’ or ‘wrong’. As long as this language has the resources to attempt comparisons between the different values its users appeal to - perhaps doing this well, in the eyes of the community, is part of what is involved in their thick moral concept best translated by ‘balance’ - it could sustain some degree of foundational practical deliberation. Indeed, there seems to be no barrier in principle to some people expressing their moral enquiry simply by ascending a chain of more abstract and general imperatives without using any predicative vocabulary of evaluation at all: consider, from our earlier case, the sequence ‘Turn right! Do not walk over the ancestral graves! Respect the ancestors! Do honour!’ Each new step in the sequence comes closer to what the speaker thinks of as the basis of practical deliberation. I am not trying to say that the imperative moral language has some kind of priority over complex predicative moral language, and that the latter is ‘nothing but’ the former in an elaborate disguise, nor would I want to make the opposite point. I am only trying to draw attention to linguistic possibilities. My point is that moral enquiry is a well-defined human practice that can be expressed in different linguistic forms, and is thus identifiable apart from any particular linguistic forms.

Let us now return to Kantsberg. As Dr Evil set about his schemes, before or after they had been revealed, Kantsbergers deliberated about what to tell him. At various points, they could have lied, or told the truth. They were initially more reluctant than the Rosstovians to choose lies because of a more fundamental choice, the choice to prioritize truth-telling in their practical deliberations. They had engaged in foundational practical deliberation, or moral enquiry, and expressed their decisions using such sentences as ‘There is no right to tell lies for philanthropic reasons’. In the version of the story where they accept that lying does not truly violate the categorical imperative, the process of foundational practical deliberation is ongoing, and their use of ‘right’ and related terms alters accordingly, in line with a basic disposition to use those terms to express their moral enquiries. They can even confront more radical differences in approaches to practical deliberation, by disputing directly with a Rosstovian or Utilitopian. They have no trouble understanding comments such as ‘It may violate the categorical imperative, but it’s right’ and indeed are able to anticipate and answer the arguments others give in support of them. They know what it is like to be convinced of a moral position, or committed to a stance on the basis of practical deliberation, and have a rough sense of what it would take for them to become convinced of a different one, and know just how they would express themselves in such an event. The most basic disposition, then, is to use certain vocabulary as a vehicle for moral enquiry.

Having identified the important dispositions in the Kantsbergers use of moral vocabulary, there remains the question of how exactly this vocabulary is to be interpreted, given these dispositions. Let’s focus on ‘is wrong’. By hypothesis, the Kanstbergers use ‘is wrong’ to express their negative moral judgements, to offer a fundamental reason not to adopt a course of action, and this is so however they exercise their moral judgement, whether in agreement with Kant or not. Across all the normal cases in which Kantsbergers do this, on what interpretation of ‘is wrong’ do they know the most? As previously indicated, I think this is a substantive meta-ethical question which a theory of reference alone cannot settle. What, if anything, do we come to learn when we engage in moral enquiry? There are various specific actions we might take during it: consulting intuitions or moral sentiments, seeking consistency, attempting to generalize or universalize principles, imaginatively projecting ourselves into perspective and situations, cultivating specific virtues, immersing ourselves in disciplined ways of life with associated spiritual practices, and so on. What do we gain from all this?

On what I have called a ‘high view’, we gain knowledge of mind-independent moral properties, or facts about human flourishing. The path to this knowledge is open to Kantsbergers, Rosstovians, and Utilitopians alike, and so the interpretation that optimizes everyone's dispositions to know is that on which 'is wrong' refers to the mind-independent property of wrongness. For even if they are often mistaken about what is wrong in their own and nearby circumstances, the tendency of moral enquiry to reach knowledge means that across the whole spread of normal cases, they will tend to learn more about the mind-independent property of wrongness, and most of their uses of 'is wrong' will express knowledge if it refers to that property.

On a lower view, we might think that we only ever learn about social convention or personal preference. In which case, a variety of options open up. At one extreme, my theory can support a potential bridge from a basic social convention view to something more like a facts about human flourishing view. *If* there is some set of basic social conventions that are uniquely stable for human (and potentially other rational?) beings, then moral enquiry would yield knowledge about those conventions across normal cases. 'Is wrong' would, in everyone's mouth, be equivalent to 'violates the most stable set of basic social conventions'.

Of course, we might take the lower view on which there is no uniquely stable set of conventions. In which, reference failure looms. As previously indicated, I wish to avoid reference failure as far as possible, especially where the appearance of meaningfulness is so compelling (see the discussion in 5.A.II and 5.C.III). One option would be to move down the dispositional ladder: since the most explanatorily basic disposition fails to yield a determinate reference, we will instead interpret the vocabulary according to the most explanatorily basic disposition which *does* yield a determinate reference. Plausibly, this implies that for the Kantsbergers 'is wrong' is indeed equivalent to 'violates the categorical imperative'. Without any anchor out in the world, it turns out that there is no referential stability for moral language. Kantsbergers and Utilitopians fail to contradict one another. Though perhaps this possibility can be made more palatable with the reflection that they might still succeed in commending to one another different approaches to foundational practical deliberation.

At this point, one might wonder whether this commendation of approaches to practical deliberation, rather than any attempt to describe the world, is not really the true function of moral language. As a whole, my theory of interpretation is designed to be realist and referentialist. The function of this thesis is to set out this realist and referentialist theory as well as I can. It is, however, perfectly true that there is more to language and thought than referentialist realism. It is beyond the scope of the thesis fully to integrate the referentialist machinery of my theory with all the various non-referential roles of language. I will record, however, that I am open to the possibility of supplying non-referential interpretations in instances where my machinery throws up a reference failure, in order to avoid total absences of meaning. If 'low' meta-ethicists wish to supplement my overall theory of interpretation with a suitable account of moral vocabulary to fill the hole their and my theories jointly generate, I am happy for them to do so.

For my part, I prefer a course of mysterianism about moral language on the low view. I do not know what the correct interpretation of moral language might be, given the low assumptions, or whether the correct interpretation is stable across profound moral disagreement. I also think there are good structural reasons for this ignorance. For what matters here are not just the meta-ethical facts, but further, higher-order facts. If we are not optimistic about moral enquiry generating knowledge, should we be optimistic about *meta-ethical* enquiry generating knowledge? Should we expect human beings to know the low meta-ethical truth across most normal cases? How will knowledge of the low meta-ethical truth affect their use of moral language? If we are optimists, then the way knowledge of this truth affects use of moral language will be crucial to interpretation, since most of the normal cases will be cases where the truth is known. But it strikes me that our basis for speculating on these questions is pretty flimsy. If someone has some confident answers, I would like to hear them.

For now, though, I suggest that we simply are not in a good enough position to interpret moral language given the low assumptions.

On the dispositions to know view, then, moral terms and concepts are referentially stable if moral enquiry is fruitful. As long as the explanatorily basic dispositions to use a given concept are grounded in moral evaluation, then that concept will express the most knowledge across normal cases if it refers to genuine moral properties, even allowing significant error in the circumstances of use. Kantsbergers may confidently assert 'All lies are wrong', but only because they happen to accept a (let us suppose) false moral theory. Given that they are trying to engage in moral enquiry as best they can, and such enquiry tends to produce knowledge, then in most normal cases they will dissent from 'All lies are wrong', and instead utter sentences that they are in a position to know if their word 'wrong' refers to the special, mind-independent property of moral wrongness. So that is what the word refers to in Kantsberg English, prevalence of error notwithstanding.

Things are more complicated on what I call the 'low' view, but, as explained, I think that referential stability and optimism about moral enquiry are naturally complementary components of a 'high' view of morality. At the very least, my theory does not straightforwardly predict that Kanstbergers and Utilitopians are talking past one another, though that is a possibility I do consider. Those wedded to consistency of meaning despite moral disagreement even on a low view may exploit the apparent failure of reference to supply a non-referential interpretation of moral language. Meanwhile, I believe that we simply do not know enough about the relevant facts to set about interpretation at all, on 'low' assumptions.

Overall, therefore, it seems that knowledge-based theories of content are adequate of themselves, and belief alone without action is a sufficient basis for interpretation. In the general case, action is a great aid to the concrete practice of interpretation, or how *we* decide what the facts of meaning are. Our primary interest, however, is how *the facts* decide what the facts of meaning are, and the facts about dispositions to know are enough for that purpose. They are even enough to secure stability for our moral vocabulary, given certain assumptions, since the most explanatorily basic dispositions to use such vocabulary are found in the process of foundational practical deliberation. Thus interpreting moral vocabulary requires us to consider all cases where that vocabulary is used for moral enquiry, and differences in specific judgements and even general principles endorsed are not decisive. Supposing for the sake of the argument that there are moral realities to be discovered, this means that moral language refers to those moral realities, and not to whatever its users happen to take for moral reality. Things are more complicated if there are no such realities, but in that case complication is no more than we should expect.

Conclusion

In this thesis I have attempted to examine the Interpretationist tradition in meta-semantics, and develop my own Interpretationist theory of content. I believe that I have done enough to show that mine is a viable theory, worthy of serious consideration. Naturally, however, there remain many ways for this theory to be developed, and outstanding issues for it to address. I will close by reviewing some particularly significant directions for further enquiry.

One is whether my theory predicts reference-magnetism: the view that some properties, being more natural than others, are thereby more eligible candidates for reference[125]. In a basic case, my use of 'green' would refer to green objects, rather than objects that are grue - either green until 2025 or blue thereafter - despite all the green objects I have seen also being grue objects, because greenness is a natural property, with the objects that in share it being objectively similar in a way that the grue objects are not.[126] Reference magnetism is appealing because it offers a clear and general solution to worries about the underdetermination of reference, and can also be applied to illuminate special contested cases: Brian Weatherson advocates for the justified true belief theory of knowledge partly on the basis that justified true belief is more natural, and thus a better candidate for the referent of 'knowledge', than many alternatives.[127]

A promising answer to this question that it would be good to defend in detail is that the optimizing dispositions to know theory does indeed predict reference magnetism, since induction over more natural properties tends to yield more knowledge than induction over less natural properties. On the 31st December 2024, I can know that any emerald I see the next day will be green, but I cannot know that it will be grue. An interpretation on which an agent refers to more

[125] See Lewis 1983.
[126] The grue puzzles originates in Goodman 1995.
[127] Weatherson 2003.

natural properties is more likely to be correct, other things being equal, because on it the agent will be better disposed to know through induction. Williams explains exactly how his rationality maximization theory predicts reference magnetism, and I believe a similar account can be supplied for mine.[128]

I would also be interested in applying the theory to debates about the methodology of metaphysics. I have already discussed how such debates inform the background of Williamson's view, and likewise they are what first inspired my own work in the area. Some authors, such as Eli Hirsch[129] and Amie Thomasson,[130] believe that applying something like a Principle of Charity can deflate metaphysical debates entirely. Firstly, we record the ways in which ordinary language implicitly answers our vaunted metaphysical questions. Then, applying our Principle of Charity to the relevant parts of English usage, we derive the result that sentences answering our metaphysical questions are true, and go happily home. Others authors - including, as we have seen, Ted Sider - take a diametrically opposed approach.[131] They reject entirely the idea that ordinary language, common-sense belief, intuition, or anything similar has any role to play in deciding metaphysical questions.

Once again, I would advocate for a golden mean. The ordinary language position is importantly similar to the view of moral language as unstable. Just as we had Kantians supposedly speaking the truth in their Kantian language, so we here have, say, mereological universalists supposedly speaking the truth in their universalist language. Just as in the parallel cases, my theory does not predict such a simple solution. Instead, we would need to identify what are the explanatorily basic dispositions to use the relevant vocabulary, and establish what knowledge is acquired following those dispositions across normal cases.

[128] Williams 2020 Chapters 3-4.
[129] See Hirsch 2008.
[130] See Thomasson 2015.
[131] Sider 2013.

This is a much more involved process, requiring something much more like metaphysics as it is generally practiced, than what ordinary language critics offer.

Meanwhile, I am satisfied that Williamson's argument against the alternative extreme is strong, and works well enough with my own view. The nature of content is such that beliefs quite generally tend to be knowledge. Thus we are not implying any implausible epistemic luck on our part by proffering 'intuitive' beliefs, such as the belief that there are mountains in northern Italy, as evidence in metaphysical enquiry. Dismissing such appeals on general grounds, as mere prejudice, is dialectically inappropriate. The critic had better give some more detailed and specific reasons for denying either that there are mountains in northern Italy, or that we know about such mountains. The potential availability in principle of such reasons, broad assent to 'There are mountains in northern Italy' notwithstanding, is of course the crucial detail ignored by the ordinary language party.

Another issue that ought to be addressed is one that appeared in passing both in the final chapter and the introduction. Language and thought has other roles besides describing reality, but my theory is entirely framed around the descriptive role. Some account of how words and thoughts end up doing the rest of the things that they do is owed. As previously admitted, I do not yet have an answer, though I note that any theory of meaning to which the descriptive role is central, as are most of those in which I am in direct conversation, share a similar burden. There are not only the obvious cases to consider, such as fiction, humour, metaphor, and so on, but also more philosophically fraught cases, as we saw when discussing morality. It would be useful to develop a strategy for deciding cases where it is contested whether the descriptive role is being performed, or some other role is being performed instead.

Finally, I have consistently treated thought and language as if they were equivalent. Of course, this is another idealizing assumption. Language is public, shared between many agents, and so interpreting any given agent properly as the speaker of a public language requires more than just considering that agent's own dispositions to believe. So, how should public languages be interpreted? Presumably through some kind of aggregation. The obvious solution would be to optimize dispositions to know across all users of the language, though whether that is the best solution, and how it would work in finer detail, requires further investigation. Once a framework for true linguistic interpretation is in place, moreover, there arises the question of how use of a public language affects the content of an agent's thoughts. It is natural to assume, and in this thesis I have often assumed, that many of our concepts are directly equivalent in content to words of our first language (and possibly others). I expect that this is the result of dispositions to believe on the basis of testimony, though again the detail wants working out further.

The goal of this thesis has been to develop a theory of content. After reviewing the development of the Principle of Charity in the 20th century, I focused on two recent Interpretationist projects: the rationality maximization of Robert Williams, and the knowledge maximization of Timothy Williamson. I judged that a theory which combined the strengths of both of these would succeed. So I proposed a theory of content based on optimizing dispositions to know. The correct interpretation of an agent, in my view, is that on which they are best disposed to know. By combining the virtues of both earlier theories, it mitigates the flaws of each. This theory delivers plausible verdicts in some interesting edge cases, and it does not suffer from focusing entirely on belief without accounting directly for action. Optimizing dispositions to know is a strong theory of content.

Bibliography

Anderson, Charity (2017). ‘Between Probability and Certainty: What Justifies Belief By Martin Smith’. *Analysis* vol. 77 (3):670-672.

Aristotle, W.K.C. Guthrie. *On the Heavens*. 1939, Harvard University Press: Cambridge, MA.

Bird, Alexander (2007). ‘Justified Judging’. *Philosophy and Phenomenological Research*, vol . 74 (1): 81-110.

Block, Ned (1981). ‘Psychologism and Behaviorism’. *The Philosophical Review*, vol. 90, no. 1, 1981: 5–43.

Brandom, Robert (1994). *Making It Explicit*. Harvard University Press: Cambridge, MA.

Burge, Tyler (2003). ‘Perceptual Entitlement’. *Philosophy and Phenomenological Research*, 67 (3), 503-548.

Carter, Sam & Goldstein, Simon (2021). ‘The Normality of Error’. *Philosophical Studies* 178 (8):2509-2533.

Chalmers, David (2005). ‘The Matrix as Metaphysics’, in Christopher Grau (ed.), *Philosophers Explore the Matrix* (2005). Oxford University Press: Oxford, UK.

Davidson, Donald (1973). ‘Radical Interpretation’, reprinted in *Inquiries into Truth and Interpretation* (2001). Oxford University Press: Oxford, UK.

-(1976) ‘Thought and Talk’, reprinted in Inquiries into Truth and Interpretation (2001). Oxford University Press: Oxford, UK.

-(1976) ‘Reply to Foster’, reprinted in *Inquiries into Truth and Interpretation* (2001). Oxford University Press: Oxford, UK.

-(1983) ‘The Coherence Theory of Truth and Knowledge, in Ernest Lepore (ed) *Truth and Intepretation: Perspectives on the Philosophy of Donald Davidson*. Blackwell Publishing: New York City, NY.

-(1987). ‘Knowing One’s Own Mind’. *Proceedings and Addresses of the American Philosophical Association*, 61: 441–58.

Frege, Gottlob (1997). *The Frege Reader* (ed. Michael Beaney). Blackwell Publishing: Oxford, UK.

Fodor, Jerry (1975). *The Language of Thought*. Harvard University Press: Cambridge, MA.

Forster, Malcolm R. & Sober, Elliott (1994). 'How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions'. *British Journal for the Philosophy of Science*, vol 45 (1):1-35.

Glock, Hans-Johann (2008). *What Is Analytic Philosophy?* Cambridge University Press: Cambridge, UK.

Goldman, A. I. (1976).' Discrimination and Perceptual Knowledge'. *The Journal of Philosophy*, 73(20): 771–791.

Goodman, J. and B. Salow (2023). 'Epistemology Normalized'. *Philosophical Review* 132, (1):89-145.

Grandy, Richard (1983): 'Reference, Meaning, and Belief'. *The Journal of Philosophy*, 70 (14):439-452.

Hirsch, Eli (2008). 'Against Revisionary Ontology'. Reprinted in *Quantifier Variance and Realism* (2011). Oxford: Oxford University Press.

Holland, Tom (2003). *Rubicon*. Little Brown: London, UK.

Horgan, T. and M. Timmons (1991). 'New Wave Moral Realism Meets Moral Twin Earth'. *Journal of Philosophical Research* 16, 447–465.

Hume, David (2000) [1740]. *Treatise of Human Nature*. Oxford University Press: Oxford, UK.

-(2007) [1748]. *Enquiry Concerning Human Understandin*g. Cambridge University Press: Cambridge, UK.

Korsgaard, C. M. (1986). 'The Right to Lie: Kant on Dealing with Evil'. *Philosophy & Public Affairs*, 15(4), 325–349.

Lakatos, Imre (1970). ‘Falsification and the Methodology of Scientific Research Programmes’, in Imre Lakatos & Alan Musgrave (eds.), *Criticism and the Growth of Knowledge*. Cambridge University Press: Cambridge, UK.

Lassonen-Aarnio, Maria (2010): ‘Unreasonable Knowledge’. *Philosophical Perspectives* 24 (1): 1-21.

Lavin, Andrew (2019). ‘The Explanatory Link Account of Normality’. *Philosophy*, vol. 94 (4):597-619.

Lewis, David (1970). ‘How to Define Theoretical Terms’, reprinted in *Philosophical Papers vol. 1* (1983). Oxford University Press: Oxford, UK

-(1972). ‘Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy,* vol 50 (3): 249-258.

-(1974). ‘Radical Interpretation’, reprinted in *Philosophical Papers vol. 1* (1983). Oxford University Press: Oxford, UK

-(1975). ‘Languages and Language’, reprinted in *Philosophical Papers vol. 1* (1983). Oxford University Press: Oxford, UK.

-(1983). ‘New Work for a Theory of Universals’. *Australasian Journal of Philosophy* 61 (4): 343-377.

.

-(1984). ‘Putnam’s Paradox’, reprinted in Papers in Metaphysics and Epistemology (1999). Cambridge University Press: Cambridge, UK.

- (1994). ‘Reduction of Mind’, reprinted in *Papers in Metaphysics and Epistemology* (1999). Cambridge University Press: Cambridge, UK.

-1996: ‘Elusive Knowledge’, reprinted in *Papers in Metaphysics and Epistemology* (1999). Cambridge University Press: Cambridge, UK.

Martin, M. G. F. (2009). ‘Reupholstering a discipline: Comments on Williamson’. *Philosophical Studies*, 145, 445–453.

McKinsey, Michael (2018). ‘Skepticism and Content Externalism’ in Edward N. Zalta (ed) *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition). URL=<https://plato.stanford.edu/archives/sum2018/entries/skepticism-content-externalism/>.

Smith, M. (2010). ‘What Else Justification Could Be’. *Noûs*, 44(1), 10–31.

McGinn,Colin (1986): ‘Radical Interpretation and Epistemology’, reprinted in *Knowledge and Reality* (1999). Oxford University Press, Oxford UK.

McGlynn, Aidan (2011). ‘Believing Things Unknown’. *Noûs* 47 (2):385-407.

-(2012a). ‘Interpretation and Knowledge Maximization’. *Philosophical Studies* 160 (3):391-405.

-(2012b). ‘Justification as ‘Would-Be’ Knowledge’. *Episteme*, 9 (4):361-376.

-Forthcoming. ‘Known Unknowns and the Limits of Knowledge’, in B. Roeber, M. Steup, J. Turri and E. Sosa (eds.), *Contemporary Debates in Epistemology*, *Volume 3*. Wiley-Blackwell.

McHugh, Connor (2011). ‘What Do We Aim at When We Aim at Belief?’. *dialectica*, 65, (3), 369-392.

Millikan, R.G. ‘In Defence of Proper Functions’. *Philosophy of Science* 56(2): 288–302.

- (1996). ‘On Swampkinds’. *Mind & Language,* 11: 103-117.

Nagel, Jennifer (2023). *Recognising Knowledge*. The John Locke Lectures: Oxford University. Available online at https://www.philosophy.ox.ac.uk/john-locke-lectures (accessed 13/07/2023).

Plantinga, Alvin (1993). *Warrant and Proper Function.* Oxford University Press: Oxford, UK.

Putnam, Hilary (1981). ‘Brain in a Vat’ in *Reason, Truth, and History* (1981). Cambridge University Press: Cambridge, UK.

Quine, Willard van Ormand (1960). *Word and Object*. The MIT Press: Cambridge, MA.

Sider, Theodore (2013). ‘Against Parthood’, in Karen Bennet and Dean W. Zimmerman (eds) *Oxford Studies in Metaphysics*, *Volume 8,* 2013. Oxford University Press: Oxford, UK.

Sinhababu, Neil (2017). *Humean Nature: How desire explains action, thought, and feeling*. Oxford University Press: Oxford, UK.

Thomasson, Amie (2014). *Ontology Made Easy*. Oxford University Press: Oxford, UK.

Weatherson, Brian (2003). ‘What Good Are Counter-Examples?’ *Philosophical Studies* 115 (1):1-31.

Williams, JRG (2020) *The Metaphysics of Representation.* Oxford University Press: Oxford, UK.

Williamson, Timothy (20000). *Knowledge and Its Limits.* Oxford University Press: Oxford, UK.

- (2007). *The Philosophy of Philosophy.* Blackwell Publishing: Oxford, UK.

- (2009). ‘Replies to Ichikawa, Martin and Weinberg’. *Philosophical Studies*, 145, 465–476.

-(2015) (forthcoming): ‘Justifications, Excuses, and Sceptical Scenarios’. In Fabian Dorsch Julient Dunant (eds) *The New Evil Demon*, forthcoming. Oxford University Press.

-Forthcoming. *Overfitting and Heuristics in Philosophy*. Oxford University Press: Oxford, UK.

Wilson, J. N. (1959). ‘Substances Without Substrata’. *Review of Metaphysics*, 12 (4): 521-539.