The Dissertation Committee for Bryan Randall Register Certifies that this is the approved version of the following dissertation:

Donald Davidson and Moral Realism

Committee:

Daniel Bonevac, Supervisor

Nicholas Asher

Joshua Dever

Cory Juhl

Ernest D. Sosa

Donald Davidson and Moral Realism

by

Bryan Randall Register, B.A.; B.S.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December, 2006

Acknowledgements

I want to acknowledge the intellectual guidance of my committee members and, most important, the unflagging support of my wife, Sarah.

Donald Davidson and Moral Realism

Publication No._____

Bryan Randall Register, PhD

The University of Texas at Austin, 2006

Supervisor: Daniel Bonevac

My thesis is that Donald Davidson's approaches in philosophy of mind, philosophy of language, semantics, and metaphysics provide the basis for an attractive version of moral realism. To show the philosophical interest of this thesis, I first contend that Davidson's approaches are themselves attractive or at least plausible. Davidson's fundamental views imply externalism about the content of mental attitudes and utterances, as well as a modest holism about the attitudes; furthermore, they entail an anti-sceptical argument concluding that coherence is an adequate test of truth, and that there can be no alternative conceptual schemes. Moral realism must account for the practical nature of moral beliefs, as well as showing that moral beliefs justified by coherence can be true. I contend that the holism of the attitudes accounts for the practical nature of moral belief. Finally, the general anti-sceptical and anti-relativist arguments work together to defeat scepticism about coherent moral beliefs.

Table of Contents

0 Introduction

The thesis of this dissertation is that the materials provided by Donald Davidson's philosophy of language provide the basis for an attractive version of moral realism to be called Hermeneutical Moral Realism. I explain here why this thesis is interesting, and outline the discussion to follow.

Moral realism is a position in meta-ethics claiming that some moral utterances are true or false, and that, among those, some are true. This position is perennial. It has never lacked for supporters or critics, and its supporters have never lacked for arguments for their view. I merely continue a long tradition. Of course, moral realists typically appeal to theses in metaphysics to support their meta-ethical view, and again, I follow a long tradition. What's different is that I find myself attracted to the core theses, and many of the supporting arguments and extended implications and articulations, of Donald Davidson's philosophy. This position has been employed as a basis for meta-ethical discussions very rarely, and never very interestingly. If I wanted to use these theses, I felt a strong need to show that they were at least plausible, or, as I see them, very attractive. For what is the interest of a defense of moral realism on the basis of unattractive or implausible metaphysics?

Because of the nature of my thesis, this dissertation has had to be quite long. To give my defense of moral realism interest, I had to offer a justification for fundamental Davidsonian claims. That occupies me in the first two chapters. I had to apply those claims in epistemology. That occupies me in the third chapter. And then I had to apply them in meta-ethics. That occupies me in the fourth and fifth chapters.

I should remark on two principles, one substantive and one methodological, that I find that I've adhered to throughout, though not on a principled basis. The first is that some concepts are too complicated to articulate. The concept of truth, and the thin moral concepts of goodness and rightness, are beyond our powers of articulation. We can never

offer theories of these concepts in the sense of a narrow cluster of claims from which all other claims about them can be derived. For these concepts are equal in extent to the mind itself, and no mind can understand itself in detail.

The second is that one useful way of proceeding in philosophy is by way of consideration of close alternative theories to one's own. Considering a near competitor gives insight into one's own view; rejecting the major near competitors lends strong support to one's own view. I don't mean to suggest solipsistic "opposition research." I've learned a good deal from every view I examine.

Let me outline the discussion. In the first chapter, I offer Davidson's approach to truth. For Davidson, the concept of truth is substantive and involves worldly connection, but not articulable in any form of "theory of truth." As I put it, the concept of truth outstrips our capacity to express it.

Davidson's approach finds its roots in Tarski's semantic theory of truth. But Tarski's theory has also inspired minimalism treatments. I discuss one such approach, Paul Horwich's, and find it wanting. That treatment gives me a better sense for the substance of truth, and also defeats one of the nearest competitor theories. I take on Horwich's use theory of meaning *en passant*. A use theory of meaning is attractive in various ways, and Horwich's comes attached neatly to his minimalism. Considering the pair as such gave me more insight into either theory than considering each alone would have done.

Tarki's (and hence Davidson's) approach is semantic in the sense that it makes truth depend on a worldly connection. That approach, likewise, is very close to traditional correspondence theories. Again, I take on an alternative theory; in this case, Russell's. I discuss the development of Davidson's objection to correspondence theories, and show why his final work presented a consideration which, though traditional, was new in Davidson's work and which combines with his older approaches to refute the traditional correspondence theory.

Davidson's own approach is humble. He does not offer a theory of truth. Rather, all other semantic concepts are to be understood with reference to truth, which is a basic, never-to-be-articulated concept. However, we can say some interesting things about truth. The formalism of a Tarskian theory is itself revealing enough to account for such phenomena as the unity of the proposition. An assertion's or belief's content is its truth-condition, which is worldly (though not to be confused with a "fact," as that term has come to be used in technical philosophy). Some such truth-conditions are events with causal powers. That fact is crucial to what will follow.

With Davidson's approach to truth clear, I can move on to the even more controversial position that meanings are truth-conditions. Chapter two is less responsive and more constructive than chapter one. I do discuss alternative approaches to meaning. The use theory retains its charms, and internalist approaches are ever-popular. I reject both. Internalism, though, is instructive because it is a penetrating solution to a real problem, the one Frege pointed out in "Sense and Reference." I begin to offer an alternative solution that, I hope, adequately treats of the intensionality of content attributions, without internalizing attributed content.

Davidson's account of meaning is externalistic and holistic. Meaning and attitudes share contents, and they are both assigned in radical interpretation; they *exist* only as a consequence of interpretation and the triangulation between interpreter, speaker, and the world.

The externalistic aspect of content is that contents are truth-conditions, external and worldly entities. Given evidential constraints on the assignment of contents to utterances and hence attitudes, nothing else could be contents. I describe the sort of constraints on interpretation that show that truth-conditions are contents, and what sort of principles interpretation must follow in order to discover meanings.

The holistic aspect of content is that attitudes have content at all only because they fit in a network of other attitudes. The intensionality of content first observed in

contact with internalism shows that we do not attribute attitudes without attention to their context.

Two other aspects of content are worth discussing. First, content is limitedly indeterminate. This is an important fact about Davidson's approach with a major bearing on its plausibility. I try to limit the scope of semantic indeterminacy, and make the remaining indeterminacy palatable. Second, content is knowable though external. This is important because a main line of response to externalistic anti-sceptical arguments like Davidson's tries to show that externalism doesn't so much refute scepticism as change its direction. Internalists might be in doubt about the world, but they know their own thoughts; externalists might know the world, but are left in doubt about their thoughts. I offer an account of self-knowledge that tries to solve the paradoxes the Davidsonian position appears to present.

In the third chapter, I turn to a pair of applications: Davidson's anti-sceptical and anti-relativistic arguments. As a prelude, I discuss Kant's treatment of similar issues; that allows me to introduce some Kantian notions that I will have occasion to call upon later. Davidson offers two anti-sceptical arguments. The first, known as the Omniscient Interpreter argument, has been widely discussed and, to my satisfaction, refuted. I diagnose the problem with the argument: its premises fail to appeal to the externalistic aspect of content. Later, Davidson offers an argument running straight from externalism and holism to the defeat of scepticism. I discuss this argument in some detail and find that it slays the sceptical dragon. While the concept of truth escapes our grasp, we can show that a coherent body of beliefs must have mainly true members. Coherentism seems to be the most plausible approach to justification, though not, of course, to truth.

I then turn briefly to the issue of relativism. Relativism is hard to articulate, and I find in the end that no version of relativism is both coherent and interesting. The coherent versions serve only to remind us that interpretation is sometimes difficult; the interesting versions make no sense on a truth-conditional account of meaning.

Next, I turn to meta-ethical issues. Moral realism is two claims: some moral utterances are true or false, and some of those are true. Various philosophers have rejected the first claim. Their thought constitutes a tradition of non-cognitivism about ethics. To defend moral realism, I discuss the prominent non-cognitivism of Simon Blackburn.

Expressivism has not been expressed with great clarity by its defenders. I begin by formulating its theses in terms drawn from Searle's illocutionary act theory. I then turn to Blackburn's justifications for expressivism. He offers arguments having to do with metaphysical parsimony, with the supervenience of the moral on the non-moral, and with moral motivation. I find that moral cognitivism is no less parsimonious than cognitivism about anything else, that expressivism has no account for the supervenience of the moral on the non-moral, and that it chooses a view about moral motivation from an inadequate set of options.

I follow on these undercutting considerations with positive arguments against the view. Expressivism has it that moral utterances are neither true nor false, but we certainly treat them as though they were: we embed moral utterances in various kinds of truth-conditional contexts. Blackburn has offered several ingenious theories for why non-cognitive discourse has every indicator of being cognitive. I show why none of them actually solve the problems.

In the final chapter, I confront moral scepticism and the nearest competitor view to Hermeneutical Moral Realism, New Wave Moral Realism. I begin by characterizing New Wave Moral Realism, which is an Aristotelian meta-ethics based in the externalism of Kripke and Putnam, Putnam's functionalism, and Rawlsian coherentism about moral theory.

I turn then to Hermeneutical Moral Realism, an original contribution to meta-ethics. I appeal to the externalistic and holistic nature of the attitudes and find some insights into moral beliefs.

First, thanks to their externalistic nature, moral beliefs' meanings are their truth-conditions. This fact gives us some hints on how to attribute moral beliefs. Further, it makes moral beliefs eligible for anti-sceptical and anti-relativistic treatment. The general arguments against scepticism and relativism should apply to moral beliefs.

Second, moral beliefs have holistic attachments to one another. This claim brings my view in contact with ethical particularism, an extremely interesting, attractive, and plausible proposal. I contend that only a modest particularism is warranted; this modest particularism is also a form of modest generalism. A more extreme form of particularism would fail to adequately respect the holistic nature of content. Particularism was the denial of principles, so I conclude that there are moral principles. However, I try to accommodate the evidence for particularism with a treatment of moral concepts as, just like the concept of truth, inarticulably complex. Moral content, I contend, flows from moral judgments to more abstract moral claims. Finally, the pursuit of Rawlsian wide reflective equilibrium represents the most plausible method for moral epistemology.

Third, moral beliefs have holistic attachments to other attitudes, notably desires and intentions. Here, I try to offer a treatment of moral motivation that adequately accounts for the evidence for the Humean theory of motivation but that is, nevertheless, Kantian in orientation. I borrow from Davidson's discussion of *akrasia*, desire, and intention, and argue that moral judgments involving thick moral concepts are holistically attached to desires and that moral judgments involving thin moral concepts are holistically attached to intentions. I belief that my approach is unique.

Finally, I briefly confront arguments for moral scepticism. Obviously, I turn to Mackie's treatment, but then I consider the Moral Twin Earth argument offered against New Wave Moral Realism. I find that the argument succeeds against the New Wave, but fails against Hermeneutical Moral Realism. What's noteworthy about this fact is that New Wave Moral Realism, like Hermeneutical Moral Realism, is based in an externalistic account of meaning and looks to a token identity theory in philosophy of mind for an

analogy to its treatment of moral properties. I diagnose why Hermeneutical Moral Realism, a Kantian view, can survive sceptical challenges that its nearest Aristotelian cousin cannot.

The strategy, then, is to prepare the ground with a broad metaphysical approach to meaning and mind, and then apply these basic insights to moral utterances and beliefs. The approach of the dissertation is very strategic, in the sense that, once the groundwork has been laid in the first chapters, I never confront an issue later on without having resources to draw on. My treatment of expressivism is fairly independent of the general strategy, but elsewhere I display little or no tactical virtuosity, but rather a sense for the place of an issue in an overall context.

1 Truth

1.0 INTRODUCTION

The basis of Davidson's philosophy, and hence my argument, is an approach to truth. Davidson's mature view is close to that of the early Wittgenstein. Wittgenstein says:

> 4.12 Propositions can represent the whole reality, but they cannot represent what they must have in common with reality in order to be able to represent it — the logical form.
>
> To be able to represent the logical form, we should have to be able to put ourselves with the propositions outside logic, that is outside the world.
>
> 5.63 I am my world. (The microcosm.) (Wittgenstein 1922, pp. 79, 151)

I can't represent the truth-making relation, for to do so, I would have to step outside myself and consider that relation from, as it were, the side. This I cannot do. Thus there is a sharp limit imposed on what can be said about truth. We can't, at bottom, say what truth is. Nevertheless, we can grasp that truth is a relational concept; truth-bearers have relations to their truth-conditions.

The nearest relatives to this approach are minimalist and correspondence theories of truth. Minimalism centers on the inarticulability of the concept of truth, but exaggerates inarticulability into contentlessness. Correspondence theory grasps that truth is a relational concept, but then tries to articulate the precise nature of the relation and the relata. That can't be done.

The correct view is not a theory of truth at all. There can't be a theory of truth, for truth is too fundamental to be articulated. Rather, the correct view treats truth as a substantive phenomenon with relation to which other phenomena can be characterized. Moreover, the correct view is willing to make claims about truth, just not claims that should be taken to reveal the nature of truth.

In the first section of this chapter, I consider minimalism about truth. I take Paul Horwich's view as my foil, since it's a justifiably prominent version of the theory. I begin by considering the nature of Tarskian theories, which inspire both minimalism and the details of Davidson's own approach — the fact that minimalism and Davidson's approach both descend from Tarski illustrates their logical closeness. I then characterize minimalism and offer various criticisms of it. I conclude by considering Horwich's use theory of meaning. The next chapter directly considers theories of meaning, but I opted to compartmentalize my discussion of Horwich.

In the second section, I consider correspondence theories. I examine a certain Russellian tradition to be found in Russell's 1912 manuscript and in recent work by Herbert Hochberg. Traditionally, Davidson has offered a classical argument known as the Slingshot as refutation of correspondence theory. I consider the merits and flaws of the Slingshot, and conclude that it provides only a substantive constraint on correspondence theories. However, in his final work, Davidson offered another classical argument against correspondence theory, the Bradley Regress. As with the Slingshot, the Bradley Regress only provides a constraint on correspondence theories. However, the two constraints working together foreclose all logical options: no correspondence theory can meet both constraints. I take this to illustrate the inarticulability of the concept of truth.

In the third section, I offer a perspective on Davidson's views about truth. Davidson's views have undergone substantial development. In earlier work, Davidson denies the existence of truth-conditions. In later work, he seems to tacitly affirm the existence of truth-conditions. I argue that Davidson's overall view makes more sense if we introduce truth-conditions; the later work is more satisfying than the earlier. Davidson's view is able to account for what the correspondence theory could not, the unity of the proposition. The main reason that truth-conditions need to be introduced is to serve as causes of beliefs. Since Davidson's theory of causation has it that causes and

effects are events, I consider whether truth-conditions can be events. I conclude that they can.

1.1 MINIMALISM ABOUT TRUTH

1.1.1 Tarski's Theory of Truth

Tarski developed his theory of truth during the dark days of positivism, when the truth predicate was widely thought to be paradoxical and metaphysical and hence indecent. To dispel these fears, Tarski sought to identify a predicate that would be adequate for all legitimate traditional uses of the truth predicate, but that would be neither paradoxical nor metaphysical. Tarski's theory has received masterful expositions[1], so I give a breezy rehearsal to hit the high points.

The paradox, of course, is the Liar. If we allow a truth (and hence a falsity) predicate into a language, then we can predicate falsity over the sentence in which the falsity predicate appears, leading to paradox. Tarski tries to solve this problem by hierarchizing languages, and asserting that, while a language may contain a truth predicate, it does not contain a truth predicate that can be applied to sentences of that very language. Attributions of truth are inherently metalinguistic.

The metaphysical aspect of truth is its irreducibility to decently physicalistic concepts. Tarski tried to solve this problem by reducing truth to other concepts that would themselves receive physicalistic reductions: satisfaction and, implicitly, translation or synonymy.

Despite the technical formality and the appearance of reform, Tarski's theory is plainly a correspondence theory intended to be in line with our ordinary intuitions about truth. In "The Concept of Truth in Formalized Languages," Tarski is clear that he "...shall be concerned exclusively with grasping the intentions which are contained in the so-called *classical* conception of truth ('true—corresponding with reality')...." (Tarski 1933,

[1] See Kirkham 1992, pp. 141-74; Soames 1999, pp. 67-97; and Field 1972, among indenumerable others. Tarski's own presentation, of course, is Tarski 1933; see also the less technical treatment in Tarski, 1944.

153) Elsewhere, he gives this insight into why he calls his theory 'semantic:' "...semantic concepts express certain relations between objects (and states of affairs) referred to in the language discussed and expressions of the language referring to those objects." (*ibid* 1936, p. 403) It seems natural, then, to say that a semantic theory of truth would be a theory that makes truth depend on a relation between objects and expressions referring to those objects. Finally and yet elsewhere, Tarski says that, "The desired definition does not aim to specify the meaning of a familiar word used to denote a novel notion; on the contrary, it aims to catch hold of the actual meaning of an old notion." (*ibid* 1944, p. 13)

Tarski demands of his theory that it be *formally correct* and *materially adequate*. Formal correctness means, more or less, solving the Liar.[2] But a definition identifying the set of true sentences, Tr, will be held to be materially adequate just in case:

> ...it has the following consequences: ...all sentences which are obtained from the expression 'x ∈ Tr if and only if p' by substituting for the symbol 'x' a structural-descriptive name of any sentence of the language in question and for the symbol 'p' the expression which forms the translation of this sentence into the metalanguage; (Tarski 1933, p. 188)

This is the famous Convention T, Tarski's criterion of material adequacy. Any theory that fails to imply each instance of the schema "'p' is true iff q," where 'q' translates object-language 'p' into the metalanguage, is materially inadequate.

Tarski wants to be able to state his theory — better, theories, since there will be at least one per language. Since, for a language with infinitely many sentences, there will be infinitely many instances of "'p' is true iff q," he needs to produce a finite substructure that has each instance of the schema as a consequence. The intuitive idea of this substructure is that there will be some small number of axioms (ideally, one) for each word of the language, and that an instance of the schema is a consequence of the axioms for the words constituting the sentence, plus finitely many rules of inference that put the axioms into inferential relations with one another. These axioms will relate words to

[2] See Etchemendy 1988, p. 54.

objects by way of the semantic concepts of reference and satisfaction. Setting aside proper names, the substructure of the theory is axioms stating the satisfaction conditions for the predicates of the language. For instance, the following might be an axiom: "for any x, x satisfies (in German) "ist weiss" iff x is white." This axiom plus some further axioms should have the implication that, "'Schnee ist weiss' is true (in German) iff snow is white," which is an instance of the schema that appears in Convention T.

There are four important points to which I want briefly to call attention. First, there is no attempt whatsoever to define truth. There is only the attempt to define truth for a given language. This is acceptable to Tarski, because the various definitions of truth for various languages will satisfy Convention T, and are hence materially adequate. Nothing more can be sought, for no sequence of characters is true or false independently of the language in which it appears, or true or false without modification. They can only be true in this language, false in that, meaningless in another.

Second, Tarski has failed to reduce truth to physicalistic notions. He has reduced truth to satisfaction, but satisfaction is not apparently a physicalistic notion.[3] Third and relatedly, in addition to satisfaction, Tarski also relies on the unreformed concept of translation. Davidson will exploit this fact to get additional use out of Tarski's formal structure, but someone like Field or Quine would no doubt object to the tacit reliance on a non-physicalistic, intensional concept. But the use of translation rather than mere material equivalence is important. That p iff q is insufficient for the correct embedding of a structural description of 'p,' along with q, in an instance of the schema from Convention T. "'p' is true iff q," where 'q' does not translate 'p,' but is merely materially equivalent to 'p,' would misstate the reason why 'p' was true (if true at all).

Fourth, Tarski's theory is obviously realistic, since a semantic theory relates words to objects. But it does not invoke new ontology. The truth conditions of an object-

[3] Hartry Field (Field 1972) objects to this feature of Tarski's theory, noting that Tarski has failed to make the notion of truth physicalistically acceptable. I agree that Tarski has not reduced truth to the physical, but I don't require that a concept be a member of the physical family of concepts to be acceptable.

language sentence are stated in a metalanguage sentence that makes no reference to anything that wasn't referred to in the object language sentence. Hence, stating truth conditions does not require the introduction of particulars, properties, facts, logical forms, or any of the other paraphernalia of correspondence theories (unless the object language sentence itself introduced these items). I believe that the theory thereby achieves the real aim of the correspondence theorist — showing that and how truth depends on a relation between the mind and language on the one hand, and the world on the other — without indulging in unhelpful metaphysical speculation.

Beyond whatever substance it's lent when the debts owed on the concepts of satisfaction and translation are paid off, the theory seems quite thin. We aren't told what truth in a language is, only which things are true in a given language and why; we're told nothing at all about truth. It's unclear whether truth is a substantive property with explanatory power. Truth isn't obviously related to meaning, the semantic concept *par excellence*. Some, like correspondence theorists, might think that these are deficiencies, but others would disagree. A minimalist would disagree; the minimalist revels in the very desiccation that might make the theory seem inadequate. Davidson would also disagree, though for a very different reason: he thinks that, while truth is connected to meaning, is a substantive property with explanatory power, and that we have an inter-linguistic notion of truth, none of this can be put into theory. Truth is too basic or foundational (not: too simple or elementary) a concept to have light shed on it by a theory; to wax poetical, truth is the light we shine on other concepts. In the rest of section 1.1, I consider minimalism and some related ideas.

1.1.2 Horwich's Minimalism

In various places, Horwich characterizes minimalism in different ways. For instance:

> The minimalist picture of truth has three principal components: first, an account of the utility of truth (namely, to enable the explicit formulation of schematic

> generalizations); second, an account of the concept of truth (namely, that 'true' is implicitly defined by the equivalence schema); and third, an account of the nature of truth (namely, that truth has no underlying nature, and that the explanatorily basic facts about it are instances of the equivalence schema). (Horwich 1990/1998, 145)

Glossing this will do for some presentation of Horwich's theory. The first component is obviously misstated. The concept of truth, or the truth predicate, rather than a property of truth, is that the utility of which must be accounted for.

The concept of truth has as its utility that it permits us to formulate schematic generalizations explicitly. Horwich has in mind sentences like, "Everything Einstein said was true." That this is an explicit generalization is obvious. But what makes it schematic? Imagine trying to get across the point of that sentence without using the truth predicate. You couldn't do it: "For anything, if Einstein said it, then it." The nearest we can come to saying it is to give the following schema, which is, strictly speaking, nonsense: "For any p, if Einstein said that p, then p." The thing about this schema that makes it nonsense is that 'p' shows up both in a 'that'-clause, and also as one of the arguments of the if-then truth-function. So 'p' is both a noun and a sentence. The schema is an extra-linguistic guide to meaningful sentences, but it isn't itself a meaningful sentence. The role of the truth predicate is to allow us to give various instances of these sorts of schemas explicitly, in a language, rather than implicitly, by picking up the right vibe from a meaningless schema. The truth predicate allows us to treat 'p' consistently as a noun, because we predicate being said by Einstein, and then being true, of it, without ever treating it as a sentence.

Why should we believe the first component? According to Horwich, the only other method we have available for forming the sort of generalizations we can form with the truth predicate is substitutional quantification, which is inordinately complicated or requires a prior explication of truth; further, the truth predicate doesn't appear to have any other utility. Since substitutional quantification seems fishy to me, too, I accept the first point. But the second point is substantive and many of the objections to minimalism turn

on it. It might be argued that we need the truth predicate to help us state and account for certain general laws concerning truth; for instance, that if something is true, then it is good to believe it; that if you're justified in believing something, then it's probably true; that we might want the truth predicate to attribute a substantive property to sentences.[4]

The second component is that the truth predicate is "implicitly defined by the equivalence schema." The equivalence schema is: "The proposition *that p* is true iff *p*." (*ibid*, 136) This seems problematic. The schema is not a sentence; it's not even meaningful. How can something that is meaningless define — give the meaning of — anything? Perhaps the meaninglessness of the schema is what accounts for the fact that the way it defines the truth predicate is *implicit*. But to define something implicitly would be to imply (perhaps with confederates) the sentences in which it is true. A meaningless schema that is neither true nor false can't imply anything. Horwich is speaking imprecisely in this instance. He's perhaps clearer when he says that,

> Because it contains no more than what is expressed by uncontroversial *instances* of the equivalence schema,
>
> (E) it is true *that p* if and only if *p*,
>
> I shall call my theory of truth '*the minimal theory*.'" (*ibid*, 6; italics prior to 'E' mine)

Here, Horwich suggests that the theory of truth expresses, not the schema, but its instances, such as "It is true that cats chase birds if, and only if, cats chase birds." That would make more sense. The schema is meaningless, but the instances are meaningful; so the series of instances seem more suitable as a definition than the schema itself.

Unfortunately, the theory/implicit definition is now somewhat cumbersome. Since each instance of the schema will be a part of the theory, and there are infinitely many such instances (as there are infinitely many truth bearers to be embedded in instances of the schema), the theory will be infinitely long. Further, there are no deeper axioms to the

[4] I veer between speaking of a property of truth and a predicate. Since I don't accept properties, resolve in favor of 'truth' being a substantive *predicate*.

theory than the instances of the schema; unlike Tarski, who gives finitely many axioms (stating satisfaction conditions for words) that will imply infinitely many sentences satisfying Convention T, Horwich has a theory with infinitely many axioms.

One should have a further qualm. It isn't strictly speaking true that "It is true that cats chase birds if, and only if, cats chase birds." Consider the sentence embedded, non-truth-functionally, within the antecedent of that conditional: "cats chase birds". Whether that sentence is true if and only if cats chase birds depends not just on how things stand with cats and birds, but also on how things stand with the words 'cats,' 'chase,' and 'birds' and the grammar of their language. If those are words of English, then the instance of the schema is true; if not, then the instance of the schema might be false. Recall that Tarski never tried to define truth because he was working with sentences, and sentences have their truth-conditions only relative to a language; hence, Tarski only defines truth relative to languages. Horwich is more ambitious. With his instances of the equivalence schema, he is trying to define truth, without modification. Hence Horwich's primary truth-bearers must not be in any language.

Why should we believe the second component? Horwich presents a two-premise argument:

> 1) ...the facts in virtue of which we mean what we do by 'true' are those that best account for our use of the term.
>
> 2) ...our use of the term is best explained by our acceptance of the equivalence schema. (*ibid*, 145)

Premise 1 is supported by a use theory of meaning. Use is, or determines, meaning; the facts that determine use, then, determine meaning. I discuss the use theory of meaning in 1.1.5. Premise 2 says that the fact that determines our use of the truth predicate is our acceptance of the equivalence schema. This premise is supposed to be supported by there being no other uses of the truth predicate than those 'explained' by the equivalence schema. But this is troublesome. The equivalence schema is a schema, not a sentence. It isn't even meaningful, much less true. So how can we accept it, and how can it explain

anything? Again, we need to replace the equivalence schema with its instances. Our use of 'true' is best explained by our acceptance of the instances of the equivalence schema. It's unclear why the instances of the equivalence schema are not uses of 'true' in need of explanation, though I suspect that the explanation will be with reference to the fact that the instances of the truth schema are necessary truths.

Why should we believe the third component, the claim that truth has no underlying nature? The most direct defense of the third component is with reference to the first and second components. But the statement of the third component leaves open the question whether truth is *itself* a nature, or property, but one that is basic (and hence not underlain by any other). Horwich accepts that truth is a property, albeit a very thin one: "...the truth predicate must indeed be rendered in logic as a predicate. Thus there is a perfectly legitimate, weak conception of property according to which minimalism implies that truth certainly is one." (*ibid*, 142) But according to stronger conceptions of properties, truth is not one: truth gets no physicalistic reduction, for instance; it has no explanatory power. Replying to the suggestion that truth is substantive but irreducible, Horwich argues:

> ...suppose that a concept of 'truth' ...is introduced by means of the stipulation that it will apply to the proposition *that snow is white* if and only if snow is white, to the proposition *that* $E=mc^2$ if and only if $E=mc^2$, and so on. Then it would seem to be consistent with our intuitive conception of 'real nature' and of 'property constitution' that the 'truth' of the proposition that snow is white consists in snow being white, that the 'truth' of the proposition that $E=mc^2$ simply consists in E being equal to mc^2, etc. — which will imply that 'truth' *as such* has no real nature. (*ibid*, 144)

The idea seems to be this. Since each truth-bearer will be true for a unique reason, if being true were the reason a thing were true, then being true would be unique for each sentence. No two instances of truth would be instances of the same thing. Or, to put the point more nominalistically, the concept of truth would be equivocal: to attribute truth to one truth-bearer is to say something different about it than one says of any other truth-bearer to which one might attribute truth.

The early Plato vibe of my gloss, "being true is the reason a thing is true," might suggest the problem. Being true is not the reason something is true, any more than piety is the reason someone is pious. Euthyphro is pious because he prosecutes his father for wrongdoing; Socrates is pious because he goes willingly to his death. The fact that the accounts for their piety are different does not imply that piety is an equivocal concept. Likewise, that *that "Snow is white" is true* is accounted for with reference to snow's being white, while *that "Grass is green" is true* is accounted for with reference to grass's being green, doesn't imply that the truth predicate is equivocal, that is, that it attributes something different to "Grass is green" than it does to "Snow is white." Horwich's argument consists chiefly in the demand that, for a concept to be univocal, the explanations for its applications must be identical; that each thing that satisfies a predicate must satisfy it for the same reason. This demand is unjustified.

Why might Horwich have gone wrong in this way? Consider multiple applications of a predicate like "is red." It's easy to see why accounting for each application of this predicate might be similar. We can (perhaps) give a reductive definition of the predicate, with reference to wavelengths of light. Whenever one is asked why something is red, one can always give the same answer: "Because it reflects light of such-and-such a wavelength." That different applications of the predicate get the same account is a marker of reduction, on a certain (early Plato) notion of what it is to account for something. If Horwich assumes that truth is either minimal, or else reductively definable, then his argument would show that it is minimal because it shows why it is not reductively definable. Were it reductively definable, each account for the application of the concept would be the same, but that is not the case. However, it may be that truth is both substantive and irreducible. Recall that this argument of Horwich's appears in the context of trying to defeat the possibility that truth is both substantive and irreducible; it appears that Horwich has begged the question.

The situation now is complicated. The third component is ambiguous; it says only that truth has no underlying property, not that it is not a basic, but still substantial, property. It's plain that Horwich thinks that saying that truth has no underlying property is enough to guarantee that is it not substantial, but he is wrong. Horwich did give a bad argument for the claim that truth is not itself a substantial, basic property. So the first two components might show that truth has no underlying nature, but they're not enough to show that it is not a basic nature, and Horwich gives us no reason to accept this latter claim.

We can resolve the ambiguity of the third component in favor of what can be supported by the first two components: the claim that truth doesn't have a nature, but might yet be one. Put in nominalistic terms, this is just the claim that truth gets only a minimal definition; it doesn't get a definition that tries to reduce it. And it seems that this is probably what Horwich has in mind; elsewhere he remarks that "...truth is not susceptible to conceptual analysis and has no underlying nature." (Horwich 1995, p. 71) He seems to identify not being susceptible to conceptual analysis with having no underlying nature, and having no underlying nature with not having a nature, which means being minimal and having no explanatory power. The step from not having a nature to having no explanatory power might be sound, but the step from not having an underlying nature to not having a nature is not. Unanalyzed (and unanalyzable) concepts might serve in explanations.

Beyond the three components, Horwich gives other explanations of minimalism. For instance, he gives five dimensions along which it is possible to 'inflate' truth; part of minimalism will be explanations of why we can get along without inflating truth along any of these dimensions.

The first such dimension is compositionality. A theory of truth might try to reduce truth, or give it underlying structure, by proposing that the truth of a truth-bearer can be accounted for with reference to the satisfaction of its parts. Such a Tarskian theory would

contain only finitely many axioms, those defining satisfaction conditions for words, and would give instances of the truth schema as theorems. Horwich says that,

> The minimalist policy is not to *deny* such principles relating truth, reference, and satisfaction, but to argue that our theory of truth should not contain them as *axioms*. Instead, they should be *derived* from a conjunction of the theory of truth and quite distinct minimalist theories of reference and satisfaction. (Horwich 1990/1998, p. 10)

The concept of a minimalist theory appears to be getting a little confused here. One was under the impression that the failure of truth to appear in explanations was a consequence of its minimal nature. But now, instances of the truth schema are going to help explain why words have their reference and satisfaction conditions. Perhaps, since the statements of these conditions are themselves minimal, the instances of the truth schema can help explain them without losing their pristine minimality.

The minimality of reference and satisfaction consists in the fact that "...*reference* [and] *satisfaction*... are just as non-naturalistic, and in need of infinite, deflationary theories, as truth is." (*ibid*. p. 117) What's not clear is why the minimalist should wish to have the concepts of reference and satisfaction at all. The role of these concepts in Tarskian theories is to serve as an axiomatic derivation base for the infinite series of theorems constituting the testable element of the T-theory for the language. The reason to have such a finite derivation base is that the human mind can't grasp an infinitely complicated theory. Since Horwich believes that we do, in fact, know infinitely complex theories, there's no role for reference or satisfaction to play in his view. That, no doubt, is why he has theories of reference and satisfaction falling out as a free consequence of his infinite theories. So Horwich is right, in a sense, that minimalism doesn't block compositionality. But the minimalist needs to explain why it is that we ever cared about compositionality at all, since his theory has no place for it other than byproduct.

The second dimension along which truth can be inflated is analysis. As we have seen, for Horwich, it is not possible to give an analysis of truth in other terms. I accept the minimalist position here, for reasons to be discussed in 1.3.

The third dimension is complexity. An inflationary theory might claim to give an analysis of truth; that such an analysis was correct would imply that truth is complex, since an analysis can only break a complex down into its parts if it is, in fact, complex. Minimalism "denies that truth, reference, or satisfaction are complex or naturalistic properties." (*ibid.* p. 11) I won't remark much on this except to point out that irreducibility or resistance to analysis does not imply simplicity. It's possible that truth can't be reduced or analyzed because it is simple. But it's possible that truth can't be analyzed because it is so enormously complex that we can't formulate the analysis. Whatever we would reduce it to escapes us. As I will contend in 1.3, that is the case; the problem with so many basic concepts, such as truth and goodness, is not that they are absolutely simple, but that they're so massively complex that we could not hope to articulate them. We tend to run together 'basic' and 'foundational,' on the one hand, with 'simple' and 'elementary,' on the other. That is a mistake.

The fourth dimension concerns the form of an inflationary theory. An inflationary theory of truth might take the form of a finite number of non-trivial statements that, when combined with other statements, allow everything about truth to be deduced. Minimalism, on the other hand, offers an infinite number of trivial statements. The lack of explanatory power of such a theory is justified by the fact that truth actually has no explanatory power, because it is minimal. Whether truth is actually devoid of explanatory power will be the focus of 1.1.4.

The fifth and final dimension concerns the connections between truth and other concepts. A theory that connects truth with, say, meaning, might be thought to inflate truth by making it central to some other philosophical issue that is plainly not susceptible to minimalist dismissals. Minimalism will try to retain the purity of truth by disconnecting it from other concepts. The crucial case is the connection of truth with meaning. Whether truth is connected with meaning can be partly shown by checking alternative accounts of meaning, ones that don't connect truth with meaning, and seeing

whether they're at all plausible. Since Horwich proposes a use theory of meaning that is intended to have precisely the effect I want to avoid — disconnecting truth from meaning — I discuss this issue in some depth in 1.1.5.

The purpose of this section has been to present minimalism, and I've done that with reference to two "bullet-point" presentations by Horwich. Obviously, the presentation has been quite critical. But I don't pretend to have refuted minimalism at this point. What I've done is present a set of difficulties; some of these might be solvable, others might not, but my goal has been to feel my way around and try to get clear on the contours of the target. The real arguments against minimalism appear in the next two sections.

1.1.3 Davidson's Critique of Minimalism

Davidson's most sustained discussion of deflationist approaches to truth appears in his papers "The Folly of Trying to Define Truth" and "Truth Rehabilitated." Across the two papers, he presents three arguments against deflationism. The first argument deals with our comprehension of the instances of the truth schema:

> The problem concerns the semantical analysis of sentences like: "The proposition that Caesar was murdered is true if and only if Caesar was murdered." The predicate "is true" requires a singular term as subject; the subject is therefore "the proposition that Caesar was murdered." Presumably it names or refers to a proposition. But then, what is the role of the sentence "Caesar was murdered" in this singular term or description? The only plausible answer is that the words "the proposition that" are a functional expression that maps whatever the following sentence names onto a proposition. In that case, the sentence itself must be a referring term.... [Possibly] in its first occurrence, the sentence names some... interesting entity. But then we do not understand the axiom, since the sentence "Caesar was murdered" is used once as a name of some interesting entity, and once as an ordinary sentence, and we have no idea how to accommodate this ambiguity in a serious semantics. (Davidson 1997, p. 10)

I mentioned this problem in the last section but set it aside. The problem is familiar: 'p' in any instance of the truth schema "it is true *that p* if and only if *p*" serves two roles: once as part of an apparently referring term ('*that p*') and once as a sentence. To allow it to serve both roles is to engage in substitutional quantification, which Horwich wants to

reject. To require it to be always a sentence is to make '*that p*' unintelligible. To require it to always be part of a name is to make it impossible to understand how it ends up as one of the truth-bearing arguments of the biconditional.

Horwich replies that this is a theory-driven objection. The problem, according to Horwich, is that Davidson's compositionalist scruples make it impossible for him to understand how 'that p' could be a name if 'p' is treated as a sentence within it:

> Davidson's... objection... is that expressions like 'The proposition that dogs bark', construed as singular terms, are unintelligible. However, this rather counterintuitive claim is entirely theory-driven: it is derived from Davidson's inability to find any account (of the sort required by his truth-theoretic paradigm) of how the referents of such expressions could be determined by the referents of their parts. (Horwich 1990/1998, p. 133)

We should probably just give up on the truth-theoretic paradigm, Horwich concludes, rather than accept that we don't understand 'The proposition that dogs bark.' Horwich seems to be going wrong in two ways. The requirement that a theory of meaning respect compositionality is not theory-driven; at least it is not driven by the truth-conditional nature of Davidson's theory of meaning. It's driven by a basic fact about language, which is that infinitely many utterances can be understood by a finite mind. The natural way to account for this fact is to give a theory of meaning that has finitely many axioms, and which accounts for the meanings of infinitely many complex utterances by reference to the meanings of their components. For any theory that accounts for compositionality, then, there must be only finitely many linguistic primitives in any language; if there were infinitely many primitives, then the language would be unlearnable. But if each and every utterance of the form 'that p' were a primitive, then a language, such as English, with infinitely many utterances of that form would be unlearnable. At a minimum, denying the compositionality of that form would require that each time we hear or use a new utterance of that form, we must learn a new linguistic primitive, which seems wrong.

To defeat this objection, Horwich needs to give a compositional account of how the meaning of 'that p' is determined by the meanings of its parts. Horwich accepts this

when he says that, "I would argue that understanding a sentence consists in nothing more than understanding its parts and appreciating how it has been constructed from them." (Horwich 1999, p. 23) Unfortunately, Horwich thinks that coming up with such an account is trivial; he continues:

> But if this is so, then compositionality is ensured, no matter what view is taken of how the meanings of words are constituted. In particular, if the meaning of a word derives from its conforming to a basic regularity of use, then the meaning of a sentence will consist in being constructed in such and such a way from primitives whose uses conform to such and such basic regularities. (*ibid.*)

Here, Horwich ignores the classical problem of accounting for the meanings of expressions in indirect contexts. Consider two expressions, p and q, which are governed by the same rules of use. Consider further a speaker who believes that p but does not believe that q. To report of this speaker that he believes that p would be to say something true, while to report of this speaker that he believes that q would be, apparently, to say something false. But, since the two expressions are governed by the same rules of use, then, on Horwich's use account of meaning, they ought to be intersubstitutable with no change in meaning. But apparently they are not. Compositionality provides an important constraint. My point here is not that the use account of meaning cannot deal with compositionality in opaque contexts. My point is only that Horwich must offer such an account. Dismissing the problem as trivial is mistaken: we might wish to regain our pre-Fregean innocence, but regaining our pre-Fregean naïveté is less attractive.

Elsewhere, Horwich actually seems to try to turn the non-compositionality of his theory into an advantage:

> ...would it not be better to pursue a Tarski-like strategy of explaining the truth of the infinitely many propositions in terms of the referents of their finitely many constituents? The answer, it seems to me, is No. The Tarski-style approach offers false hope (1) because, as is well-known, there are many kinds of proposition (e.g., statements of probability, counterfactual conditionals, etc.) whose truth we have no reason to believe can be explained on the basis of the referents of their parts; and, more importantly, (2) because such a strategy would miss those propositions that are constructed from the primitive concepts that are not expressed in our language. If *all* propositions are to be covered, then there would

> have to be axioms specifying the referents of all the infinitely many possible primitives. So the Tarskian approach would turn out to need no fewer axioms than the Minimal Theory. (Horwich 1990/1998, pp. 136-7)

There are two lines of argument. First, because we don't yet know how to make Tarskian accounts of compositionality work for some sentences, compositionality is to be rejected. Horwich seems to be insisting that compositionality is an arbitrary imposition with no basis in evidence. But it's a truism that compositionality is present in learnable languages with infinitely many possible utterances. The fact that we don't seem to be learning new singular terms whenever we hear new statements of probability or counterfactuals indicates that the meanings of these utterances are determined by their parts, the meanings of which we already knew; that's compositionality. Horwich seems to have confused requiring a theory to account for compositionality with requiring a theory to have been invented by Tarski. Second, Horwich notes that a Tarskian theory with finitely many axioms wouldn't capture every possible proposition. That's true, but whether it matters or not depends on what a proposition is. Tarskian theory is metalinguistic: we give a theory of truth *for a language*. The fact that my theory of truth for, say, German doesn't allow me to state the truth-conditions of propositions that aren't expressible in German isn't an objection to my theory. Only if the truth-bearers are non-linguistic propositions, as in Frege, does it count against a finitely axiomatized theory of truth for a language that it fails imply certain propositions. I argue below that there are no such Fregean propositions. If that's right, then the advantage Horwich claims for his theory, that it gives the truth-conditions of propositions that a language-bound Tarskian theory fails to give, won't be an advantage.

Horwich's assessment of Davidson's challenge seems off in another way. Davidson claims only that 'The proposition that dogs bark' is, *if* a singular term, incomprehensible, not that it is incomprehensible. Davidson denies, on these grounds, that it is a singular term. Horwich says that Davidson's point is counterintuitive, but it's hard to see where one might have gotten the intuition that 'The proposition that dogs bark'

is a singular term; that "intuition" seems far too heavily theory-laden, because couched in much too theoretical terms, to be a good test for a theory. Horwich explains that, "One might suspect that Davidson's attitude derives from scepticism about propositions; however he is quite explicit that this is not the objection. But in that case — if there really *are* such things — how can the expressions specifically designed to refer to them be unintelligible?" (Horwich 1999, p. 23) Horwich has gotten things backwards. If expressions of the form 'that p' refer to propositions, then they are singular terms. But if they are singular terms, then their meaningfulness cannot be consistent with compositionality. Anything the meaningfulness of which cannot be consistent with compositionality is not meaningful. By a pair of *modus tollens* inferences, Davidson arrives at the conclusion that expressions of the form 'that p' do not refer to propositions. To attack minimalism, Davidson appeals directly to the compositionality requirement, not to the additional conclusion, also based on that requirement, that propositions do not exist. Davidson's claim is not that expressions of the form 'that p' are unintelligible; it's that they would be if Horwich were right.

Not only does Davidson think that we do understand 'The proposition that dogs bark,' he offers a theory about *how* the utterance is understood, his well-known analysis of that-clauses. In that analysis, Davidson suggests that the logical structure of sentences with that-clauses is as follows: one sentence that includes 'that' as a demonstrative and says something *about* that, and another sentence that *is* that.[5] For instance:

The proposition that dogs bark is true.

...gets analyzed as:

Dogs bark.

That proposition is true.

[5] See Davidson 1968, esp. pp. 102-6.

'That' in the second proposition refers to the actual ink pattern just above itself. No intensional entities (beyond the sentence "Dogs bark") are called for.[6]

Does Davidson's analysis of that-clauses solve the problem he presents for minimalism? Horwich could give the truth schema a makeover so that it looks like this:

p

that is true iff p

...and its instances look like this:

Dogs bark.

That is true iff dogs bark.

Does that solve the problem? No: many instances of the new truth schema will be false, because the referents of the instances of 'that' are language-bound entities: sentences in languages. If we rely on homonymy between the referent of 'that' and the right side of the 'iff,' we will often go wrong, since the two homonymous utterances might be in different languages and have different truth conditions. What Horwich needs is a way for 'that' to refer to a proposition in the sense of a Fregean thought. Until such propositions are introduced, the instances of the truth schema are language-bound; the schema is more or less just disquotational.

I now move to another Davidsonian argument, one that fails for complicated reasons. Davidson argues:

> Disquotation cannot, however, pretend to give a complete account of the concept of truth, since it works only in the special case where the metalanguage contains the object language. But neither object language nor metalanguage can contain its own truth predicate. In other words, the very concept we want to explain is explicitly excluded from expression in any consistent language for which disquotation works. (Davidson 1997, pp. 10-11)

This argument works on two assumptions. First, we attribute truth to linguistic items. Second, the only possible solution to the Liar is a Tarskian, hierarchical solution in which

---

[6] Doesn't the second sentence *say* that it's about a proposition? Well, yes: it says, "That is a proposition and that is true." But all that should indicate is that 'proposition' in English doesn't necessarily refer to intensional entities other than sentences.

falsity is never predicated over a sentence of the same language of which the falsity predicate is a word. In the absence of either assumption, the argument won't go through.

Here's how it's supposed to work. If truth and falsity are attributed to linguistic items — sentences — then there must be some relation between the language of the truth predicate, the meta-language, and the language of the sentence, the object language. If the relation is identity, then the Liar emerges. If the relation is not identity, then truth can't be defined in general, but only for particular object languages. Since Horwich's minimalism defines truth through the infinite list of instances of the truth schema, Horwich's minimalism will necessarily fail.

In the absence of the first assumption, that truth-bearers are sentences, we can define truth because there need be no relationship between the language of the truth-bearers, which are not in a language, and the language of the truth-predicate. We define truth for propositions. In the absence of the second assumption, we can deal with the Liar even if the object language is the metalanguage.

Horwich rejects both assumptions. He thinks that the truth-bearers are propositions in the sense of Fregean thoughts, not linguistic items. And he believes that, while all propositions are either true or false, some sentences may be neither, which amounts to accepting a "gappiness" solution to the Liar when it's applied to linguistic items.

Let's begin with the claim that the truth-bearers are propositions. Propositions are pretty dubious entities. Horwich gives the following argument for the existence of propositions:

> Let us imagine a body of sentences characterized by their concern with a certain range of phenomena; and suppose that we have mounted an investigation into the relations of deductive entailment that hold amongst these sentences. Suppose that the results of our investigation suggest an attribution of logical forms having the implication that some of the sentences will clearly entail the existence of entities of a certain type — call them '*Ks*'. Suppose, finally, that we believe that some of those sentences express truths. Taken together, these considerations would provide a basis for thinking that things of type *K* exist....

> It is easy to see how these general conclusions will apply to the case of propositions. In the first place, we can suppose that an adequate account of the logical forms of belief attribution involves the supposition that 'that *p*' is a singular term.... In the second place, we may assume that some propositional attitude statements are certainly true.... Thirdly, we should take these assumptions to entail that there is an entity.... (*ibid* pp. 88, 89-90)

Davidson fans have to accept this argument (since it's done up so nicely in accordance with Davidson's "method of truth in metaphysics"), so I will. There are propositions. But the argument doesn't tell us what propositions are. Propositions, I would suggest, are sentences in languages. 'That *p*' is not, as Horwich suggests, a singular term. 'That' is a singular term; '*p*' is the sentence/proposition to which 'that' makes reference. Sentences are the objects of belief. That this presents me with certain problems is obvious; I'll deal with those problems later on when I come to Davidson's theory of attitude attributions.

So Horwich has not shown what he needs to show to resist Davidson's argument, which is that the truth-bearers are not linguistic. They are propositions, but propositions might well be linguistic entities. But is there any special reason to *deny* that propositions, in Horwich's Fregean sense, exist? There are reasons. I offer one Quinean argument and one Russellian argument.

The Quinean argument has to do with indeterminacy.[7] Let's take for granted (what I show in 2.4) that there is at least modest indeterminacy of meaning, in the following sense: if there is a meaning-preserving translation of an utterance from one language into another, then there are also other translations that will also apparently preserve meaning. But there's nothing more to preserving meaning than apparently preserving meaning; if the meaning that is to be preserved were something that we couldn't tell whether we had preserved, then meaning would have gone private. Indeterminacy is a consequence of the public nature of meaning.

But if meanings were propositions, then, for a translation, interpretation, or indirect attitude attribution to preserve an utterance's meaning or give an attitude's

[7] See Davidson 1968, pp. 100-1. Also, of course, Quine 1960, esp. pp. 27, 72-9.

content, one would have to pick out the meant proposition. In general, the existence of Fregean propositions would lend too great a determinacy to the states and utterances that took those propositions as their contents. Thus there are no such propositions. Horwich disagrees. He argues that indeterminacy is rampant in all walks of language, and if it doesn't lead to scepticism about anything else, it shouldn't lead to scepticism about propositions.[8]

Consider, for instance, the old heap problem. One grain of sand is not a heap; if n grains of sand are not a heap, neither are n+1 grains of sand; hence, there are no heaps of sand. Horwich rejects the second premise: "...we must allow that there is some unknown (indeed unknowable) number, *h*, such that *h* grains cannot make a heap but *h*+1 grains can. Thus we are allowing that the predicate 'is a heap' has an extension, albeit an indeterminate one. True, we could not, even in principle, discover the extension." (*ibid.* p. 81) There are two problems with this approach.

First, Horwich has confused two kinds of indeterminacy. The indeterminacy he appeals to for heaps is epistemic: anything is either a heap or it is not, but sometimes we can't tell, and in that case there's indeterminacy. But the indeterminacy of meaning goes beyond the epistemic. We can begin with merely epistemic indeterminacy: we can't tell which proposition is meant by an utterance. But with meaning, nothing is terminally hidden. Hence epistemic indeterminacy applied to meaning yields ontological indeterminacy. It's not that there's a fact of the matter but we can't know it; it's that there's no fact of the matter to be known. But if there were propositions in Horwich's sense, then there would be a fact of the matter to be known; so there are no such propositions.

Second, Horwich's account of indeterminacy appears to contradict his own use account of meaning. On the use account of meaning, the meaning of 'is a heap' is determined by its occasions of use. So whether something is a heap or not is determined by whether it occasions the use of 'is a heap.' If it occasions neither the use of that

[8] Horwich 1990/1998, p. 78-84, 94n.

predicate, nor the use of that predicate in a negative way ('is not a heap'), then it neither is nor is not a heap. It could only be a heap or not if the rules of use for the predicate told us whether to apply the predicate or not; if the rules don't say, then it's not (even unknowably) a heap or not. There can't be a pile of sand for which the rules determine whether 'is a heap' applies or not but nobody knows what the determination of the rules is. The use theory appears to undermine his notion of indeterminacy as unknowability. Absent a more serious argument against the indeterminacy of meaning, then, Horwich has no response to this argument against propositions.

The other argument against propositions is related but of an older (1905, to be precise) vintage: it's an argument at least hinted at in Russell's polemic against Frege in "On Denoting." Russell argues:

> We say, to begin with, that when *C* occurs it is the *denotation* that we are speaking about; but when '*C*' occurs, it is the *meaning*. Now the relation of meaning and denotation is not merely linguistic through the phrase: there must be a logical relation involved, which we express by saying that the meaning denotes the denotation. But the difficulty which confronts us is that we cannot succeed in *both* preserving the connexion of meaning and denotation *and* preventing them from being one and the same; also that the meaning cannot be got at except by means of denoting phrases. (Russell 1905, p. 49)

There are two lines of thought here. First, there is no way to state the relation between sense and reference ("meaning" and "denotation") that both accounts for the fact that sense determines reference and also distinguishes between sense and reference. I find this line of argument difficult to understand. One doesn't like to be dismissive toward one of the greatest philosophers of our time, but I suspect that Russell has confused himself by making a pair of errors: using a phrase in quotation marks to refer to the meaning of the phrase, but also to refer to the phrase itself; and thinking that the relation between a word's sense and its reference (i.e., the relation of reference) is the same as the relation between the word and its reference (i.e., the relation of reference). So I don't pursue this line of thought. Second, we can refer to senses ("meanings") only with reference to their

relations to linguistic items (denoting phrases), and this is problematic. Here I think Russell has it right.

How can we refer to meanings? Well, how do we refer in general? Perhaps we don't do much in the way of reference; a line of thought begun by Russell and continued by Quine would have it that our language's apparently referential apparatus is, at its deepest level, only quantifiers, variables, and predicates. But let us accept that reference occurs and even that, in at least some contexts, definite descriptions refer. Consider this apparently referential term, 'the mother of Buffy.' Now, the mother of Buffy is named Joyce. But I can manage to refer to Joyce otherwise than by using her name. I can refer to her by first identifying something with a relation to her (Buffy), and then giving a function ('the mother of') that takes one from the name following 'of' to that object's mother. The trouble with Fregean meanings is that we more or less *have* to use a method like this to refer to them. It's not apparently possible to refer to a meaning otherwise than by giving a linguistic item, and using a function that takes one from that phrase to its meaning: "the meaning of 'p'."

We don't need the non-Russellian account of the referential powers of definite descriptions; the same conclusion will follow even if definite descriptions are not referential. Let "The mother of Buffy is ill" be: there is a unique x, such that x mothers Buffy, and x is ill. We still identify x with reference to its relata, Buffy. Likewise, "The meaning of 'p' is an abstract entity" might best be read: there is a unique x, such that 'p' means x, and x is an abstract entity." Again, we identify x with reference to its relata, 'p.'

What's wrong here is that meanings are not supposed to be linguistic items. Thus their criteria of identity and individuation ought not be linguistic. But if we can identify and individuate them only with reference to their linguistic relata, then their criteria of identity and individuation will turn out to be linguistic anyway.

How does this apply to Horwich's claim about propositions? The propositions Horwich wants to make the truth-bearers amount to Fregean thoughts, which were, for

Frege, the senses of complete sentences. Thus any argument against senses should apply to Horwich's propositions. Horwich's propositions require criteria of identity and individuation. Since propositions are to be non-linguistic items, these criteria must not be linguistic. But because of their peculiar nature, propositions can only be identified and individuated with reference to their linguistic relata. The claim that propositions (in the relevant sense) exist leads to the contradiction that there are entities that can be identified otherwise than with reference to their linguistic relata, but that can only be identified with reference to their linguistic relata. Thus propositions (in the relevant sense) do not exist.

I turn to the Liar. Horwich suggests that there are four possible solutions to the paradox: 1) deny bivalence, 2) deny that truth and falsity can be predicated of sentences in a language of sentences in the same language, 3) deny that instances of 'p' in paradoxical sentences express propositions, 4) reject the paradoxical instances of the truth schema.[9] (1) and (2) are obviously quite radical solutions, to be avoided if at all possible. Given his acceptance of propositions, the obvious way to go would seem to be to adopt (3) and (4). (3) would tell us that some sentences aren't apt to be true or false, because they fail to express propositions; (4) would carry out the procedure of excluding those sentences from the truth schema. The fact that the instances of 'p' in paradoxical sentences don't express propositions would explain why we are to reject instances of the truth schema concerning 'p.' Horwich rejects (3) but adopts (4). Here's his problem with (3):

> ...for any condition *C*, one might happen to believe that the proposition meeting that condition is not true — which (since any object of belief is a proposition) would imply that 'The proposition meeting condition *C* is not true' expresses a proposition. And this will be so even if it happens to turn out that the proposition it expresses is the one meeting *C*... (Horwich 1990/1998, p. 41)

This argument seems to beg the question. Horwich assumes that everything that seems like a belief is one and hence has a proposition as an object. But this seems false. Try as I

[9] Horwich 1990/1998, p. 41.

might, I *cannot* believe that the slithy toves did gyre and gimble in the wabe. I might think that I believe that, but I'd be wrong. Likewise, even if I try to believe that the proposition meeting condition *C* is not true, there may be no proposition meeting condition *C*; that can be true even if 'the proposition meeting condition *C* is not true' might have been thought to express just precisely that proposition. Maybe it's meaningless.

The trouble with accepting (4) while rejecting (3) is that one would like an account of why the paradoxical instances of the truth schema are false. (3) gives just such an account. In place of (3), Horwich announces a set of apparently arbitrary rules to exclude the paradoxes: "(a) that the minimal theory not engender 'liar-type' contradictions; (b) that the set of excluded instances be as small as possible; and — perhaps just as important as (b) — (c) that there be a constructive specification of the excluded instances that is as simple as possible." (*ibid.* p. 42) Assuming such a specification, one would want an account of why the sentences specified to be excluded *should* be excluded, and this Horwich does not provide. It seems that by his own lights, Horwich should accept (3) and use it to explain (4) and thereby solve the Liar.

But accepting (3) involves claiming that certain sentences are problematic because they don't express propositions. Since I've argued that no sentence expresses a proposition (in the relevant sense), I'm committed to the rejection of (3) as useless. However, I imagine that there is an alternative, (3'), that would deny that instances of 'p' in paradoxical sentences are meaningful. This is a generalization of (3). (3) says what (3') says, but on the assumption that being meaningful, for a sentence, is expressing a proposition. Since I deny propositions, I deny the move that goes from (3') to the more specific (3). The job of working out some version of (3') is the job of explaining why the instances of 'p' in paradoxical sentences are meaningless. Horwich's (3) would have done that had there been propositions. Horwich's rules (a)-(c) are a pretty weak attempt to provide the basis for rejecting exactly the right instances of 'p.' However, some

specification of (3') would have to work for the truth-value gaps approach to work. Since I'm inclined to accept a truth-value gaps approach, I'm committed to there being some specification of (3') that will exclude exactly the paradoxes. I don't know of any reason Horwich couldn't help himself to the same principle, whatever it is, so I think that Horwich could, if he were so inclined, provide some justification for (4). I conclude, then, that one of the assumptions of Davidson's argument was mistaken: the hierarchical approach is not the only possible solution to the Liar because a truth-value gaps approach could work. Since Davidson's argument required both assumptions, and only one of them seems to be true, this argument is unsound.

I'll be quick with Davidson's third objection, mainly by putting it off. Davidson explains:

> Horwich recognizes that to maintain that truth has, as he says, "a certain purity", he must show that we can understand it fully in isolation from other ideas, and we can understand other ideas in isolation from it. He does not say there are no relations between the concept of truth and other concepts; only that we can understand these concepts independently....
>
> Understanding a sentence, he maintains, consists in knowing its "assertibility conditions" (or "proper use"). He grants that these conditions may include that the sentence (or utterance) be true. I confess I do not see how, if truth is an assertibility condition, and knowing the assertibility conditions *is* understanding, we can understand a sentence without having the concept of truth. (Davidson 1996, p. 33)

The objection turns on the relation of truth to other concepts. For Davidson, truth is central to meaning; for Horwich, meaning is prior to and independent of truth. But if Horwich, in giving sentence meanings, ends up giving a list of truth-conditions, then his account will, like Davidson's, make truth central to meaning. Whether Horwich makes this mistake — he does — is a matter better put off for the full discussion of Horwich's theory of meaning in section 1.1.5, so I leave off here.

1.1.4 Explanation-Oriented Critique of Minimalism

It would appear that there are general facts about truth: whether or not a sentence is true seems susceptible of explanation, and that sentences are true seems to explain further facts about them. But part of what it is to be a minimal property is to both lack explanatory power, and not to require explanation. So if truth is involved in explanation either as explanans or explananda, it is not minimal. In this section, I focus on arguments dealing with truth in explanations. I go through in two passes. First, I look at the example of the normativity of truth. Second, I look at the more general argument against minimalism offered by Anil Gupta.

We seem to desire to believe *p* only if *p* is true, and *p*'s truth seems to be important to its desirability; further, we strongly prefer to have this desire. Dummett makes the point in this famous passage:

> ...it is part of the concept of truth that we aim at making true statements.... We cannot in general suppose that we give a proper account of a concept by describing those circumstances in which we do, and those in which we do not, make use of the relevant word, by describing the *usage* of that word; we must also give an account of the *point* of the concept, explain what we use the word *for*. (Dummett 1959, pp. 2-3)

For Dummett, truth is a goal of enquiry, and this is a fact that must be accounted for on any theory of truth. Two issues emerge. Is truth a goal of enquiry? And, can any account of this fact be given if we accept minimalism?

The claim that truth is a goal of enquiry has come under attack by, unfortunately, Davidson, while engaged in debate with Rorty. Here's what Davidson has to say:

> ...truths do not come with a "mark", like the date in the corner of some photographs, which distinguishes them from falsehoods. The best we can do is test, experiment, compare, and keep an open mind.... Since it is neither visible as a target, nor recognizable when achieved, there is no point in calling truth a goal. Truth is not a value, so the "pursuit of truth" is an empty enterprise unless it means only that it is often worthwhile to increase our confidence in our beliefs, by collecting further evidence or checking our calculations. (Davidson 1997, pp. 6-7)

For Davidson, there is a tension between the objectivity and normativity of truth. If truth were normative, it would have to be visible and recognizable. But to be objective, it must be neither. Objectivity, in this context, means independence from us. Truth's objectivity would consist in the fact that whether something is true is determined by factors independent from us, our epistemic situation, what's useful for us, and so forth. Pragmatists and others opted to give up the objectivity of truth and identify it with something visible and recognizable, assertibility. But truth transcends assertibility under any specifiable conditions, unless those conditions themselves so transcend any actual situation in which anyone might find herself as to amount to infallibility. So Davidson takes the other alternative and retains the objectivity of truth, while abandoning its normativity.

The major premise of the argument is that, for something to count as a goal of an agent, that agent must be able to envision the goal while aiming at it, and recognize her achievement of it (should she in the end achieve it). This seems straightforwardly, indeed obviously, wrong as a constraint on what we should count as goals.

Consider the first constraint: envisioning. In the simple example of trying to hit a target with an arrow, it's obvious that I need not envision the target to try to hit it. I could, for instance, try to hit a certain distance to the left of some other object, knowing that the target is that distance from the other object. Here, the analogy is with the relationship between truth and justification: I can aim at truth by aiming at something with a certain relation to justification.[10] I could even use The Force or follow the Tao in my attempt to hit the target, and still count as trying to hit it.

Consider the second constraint: recognition. In the example, I need not check to see whether I have hit the target to count as having tried to hit it. Now, I do have to be able to recognize, in general, hittings of targets with arrows before I can intend to hit a target with an arrow, because in the absence of the ability, I can't take hitting a target with

[10] See section 3.2 on the tightness of this relationship between justification and truth.

an arrow as the content of any intention of mine. The proposition "I have hit a target with an arrow" isn't one that I can intend to make true, because I wouldn't understand it.

So there is something to the second constraint. But all we're required to do by the recognition constraint is to have the concept of truth. Certainly we must accept that we must grasp the concept of truth before we can take truth as a goal, but, since we *do* grasp the concept of truth, that's no objection.

At a deeper level, I want to consider an argument Putnam offered in this context.[11] Putnam argues that you can't grasp conditions of warranted assertibility without grasping conditions of truth. Consider the conditions of warranted assertibility of a sentence like, "The cat is eating." We could take more or less internalist or externalist notions of the conditions. On an internalist approach, the conditions of warrant might be: that I have a there's-a-cat-eating sort of experience. But what are the conditions of warranted assertibility of a sentence like, "I am having a there's-a-cat-eating sort of experience?" Surely, that I have a there's-a-cat-eating sort of experience. But, by no coincidence, that is the truth condition of the sentence. On an externalist notion, the conditions of warrant might be: that there's a cat eating, and perhaps I'm aware of that by some reliable mechanism. Again, the truth conditions are part of the conditions for warranted assertibility. In both cases, I have to be in a position to grasp that some sentence's truth conditions are met to be in a position to grasp that some sentence's (perhaps the same one) conditions of warranted assertibility are met. It's hard to envision a notion of warranted assertibility that can be taken as a goal across the board without also taking truth as a goal in at least many instances. It's therefore a mistake to deny that truth can be goal in favor of taking warranted assertibility as a goal; one must meet either to meet the other.

Truth can be both objective and a norm; we must be in a position to envision and recognize truth to be in a position to envision and recognize warrant, and being

[11] Putnam 1991, pp. 266-9.

envisioned and potentially recognized is not a precondition on taking something as a goal anyway. I accept the intuitive claim that truth is a goal of enquiry and reject Davidson's Rortian argument.

But what sort of explanation is necessary of the fact that truth is a goal of enquiry? I argue that, whatever the explanation is, it will have to accept that truth is a substantive property. For Horwich, we seek the truth because it's useful:

> Consider in the first place those of a person's beliefs of the form
>
> (1) <If I perform action *A* then state of affairs *S* will be realized>.
>
> The psychological role of such beliefs is to motivate the performance of *A* when *S* is desired. When this process takes place, and if the belief involved is true, then the desired result will in fact obtain. In other words, if I have belief (1) and desire *S*, then I will do *A*. But if my belief is true, then, given merely the equivalence axioms, it follows that if I do *A* then *S* will be realized. (Horwich 1990/1998, p. 44)

Horwich continues by saying that we should care about the truth of beliefs that aren't of the form (1) because of their inferential relations with beliefs that are of form (1). Let's consider some belief (1') that is of form (1), and run through the argument:

> (A) We should believe that (1') only if believing that (1') is useful.
>
> (B) *Ceteris paribus*, believing that (1') is useful only if that (1') is true.
>
> (C) *Ceteris paribus*, we should believe that (1') only if that (1') is true.

Since we can derive (C) from (A) and (B), those two claims might be thought to account for the truth of (C). Let's call this account, and all accounts that share its form but make reference to different sentences in place of (1'), instances of schema (N). The trouble with (N) is that (A) seems to be false, and (B) not provable given minimalist scruples.

What about the *ceteris paribus* clauses? I insert *ceteris paribus* in (B) to allow for the fact that sometimes a false belief might be useful, and this weakening of one of the premises will flow through to the conclusion, (C). But why doesn't (A) get a *ceteris paribus*? For (A) to get a *ceteris paribus* would be for there to be some other reason to believe a sentence other than its usefulness.

(A) is the claim that belief's point is exclusively practical. But that seems false. We like to know things just for the sake of knowing them, beyond any practical purposes. It strikes us as tragic that so many of our fellows are devoid of purely intellectual curiosity. That high-grade reflective knowledge even *has* a practical point is a fairly recent development: all men by nature desired to know, long before knowledge was power.

Consider a simple example. As it happens, William the Usurper was crowned King of England on Christmas Day, 1066. But how bad off would I be if I believed that William had been crowned on, say, All-Saint's Day? Or New Year's Eve? Is there some lesson of history that would be misapplied by anyone who had William's coronation date a bit off? Surely not. Yet it would irritate me no end to find that I'd been wrong about the date.

How much damage would rejecting (A) do to Horwich? Instances of (N) are simple hypothetical syllogisms. Whether the antecedent of the first premise can be connected in the conclusion with the consequent of the second premise depends on whether something can be found that can plausibly serve as consequent of the first premise and also antecedent of the second. That (1') is useful can't plausibly serve, because we care about truth for other reasons than its usefulness. What other reasons are there for believing things, other than that it's useful to believe them? Consider two equally useless beliefs, *p* and ~*p*, and assume that *p*. We would prefer to believe that *p* in this situation, but the two beliefs are equally useless. The only difference between the two beliefs is that one of them is true. But (A) reformed along these lines:

> (A') We should believe that (1') only if believing that (1') is useful or that (1') is true.

...isn't going to work for minimalism, since on (A'), (1')'s truth is making a difference to something else, whether we should believe it. Horwich disagrees:

> It might be thought that if truth *is* intrinsically valuable, then minimalism is in trouble, since it surely lacks the resources to explain that value. But this criticism is unjust. For the difficulty that attaches to explaining why true belief is intrinsically good is no more or less than explaining, for any other particular thing, why it is intrinsically good. The problem stems from our failure to understand the concept of intrinsic goodness, rather than from our adoption of the minimalist conception of truth. (Horwich 1990/1998, p. 62)

I suspect that Moore was right and that we're never going to understand the concept of intrinsic goodness any more than we're going to understand the concept of truth. But Horwich seems to be wrong to say that minimalism's problem with understanding the intrinsic goodness of truth has to do with intrinsic goodness. Horwich says that it's hard to understand of any "particular thing," like kindness, why it is intrinsically good. That's true, but at least with kindness there's something in the goodness of which one can believe. The problem with minimalism here is that it denies the existence of that which is commended, truth.

For minimalism, claims that seem to be about truth aren't. We only use the truth predicate to generalize. So consider again (A'), the claim we're left with when we abandon the narrowly consequentialist conception of the value of truth:

> (A') We should believe that (1') only if believing that (1') is useful or that (1') is true.

Why might we believe this? Well, the natural source for a claim like this would be a universal claim of which it is an instance:

> (UA') ∀x(we should believe *that x* only if believing *that x* is useful or *that x* is true)

But how might the minimalist discover such a claim? Likewise, what about (B), the claim that, *ceteris paribus*, believing that (1') is useful only if that (1') is true? (B) can be derived from a certain universal claim:

> (UB) *Ceteris paribus*, ∀x(believing *that x* is useful only if *that x* is true)

To purify the argument of any essential reference to truth, we would have to add the relevant equivalence axiom:

(EB) That (1') is true iff (1')

Horwich won't interpret this truism in a way that accords with compositionality, but let's assume that he makes the use theory of meaning work for the interpretation of that-clauses. Instantiating (1') for x in (UB), and then running across the biconditional in (EB), will get (B). But how would we figure out that (UB), or (UA')? Horwich's answer is amazing:

> ...it is plausible to suppose that there is a truth-preserving rule of inference that will take us from a set of premises attributing to each proposition some property, *F*, to the conclusion that all propositions have *F*. No doubt this rule is not *logically* valid, for its reliability hinges not merely on the meanings of the logical constants, but also on the nature of propositions. But it is a principle we do find plausible. (Horwich 1990/1998, p. 137)

The base for the inference to (UB), then, is the infinite list of its instances. Why must all of the instances be included? Why not use inductive or abductive reasoning on a subset of the instances to induce the generalization? To induce something about truth — that it's connected to usefulness — would be to treat truth as a substantive property, which Horwich cannot do. But note that (B) is one of the instances, and the question I asked was, How do we know that (B)? The natural answer is, because it's an instance of a generalization that we believe. But we believe the generalization, according to Horwich, because we believe *all* of the instances. There would be a vicious circle here if Horwich accepted that (B) had any support at all.

For minimalism, there are no substantive truths about truth. The truth predicate is a device of generalization. We come to the very heart of minimalism when we realize that (A') and (B) and every other claim that apparently involves truth doesn't, but is only a way of summarizing or restating some other claim, one that doesn't involve truth. Since no claim not involving truth could support (A') and (B), and there are no claims involving truth, these claims are basic and receive no justification. They must be *a priori*. But on

the face of it, a claim like (B): *Ceteris paribus*, believing that (1') is useful only if that (1') is true, is not a good candidate for *a prioricity*. The claim seems straightforwardly empirical: it's about the causal effects of believing a claim. The fact that a *ceteris paribus* clause is necessary should help us to remember why.

Recall that Horwich offered a (mistaken) account, (N), of why it is good to believe claims of form (1) only if they're true. That account turns out to have been schematic; i.e., not really an account at all. There is no general claim to account for, there are only the instances. (N) is really a pattern according to which we may create accounts of our own for each instance of (1) about which we have any interest. However, some instances of (1) will be false and yet useful to believe, or true and yet useless to believe; that's why (B) has a *ceteris paribus* clause. Yet (N) did not account for this; it was a proof of (1) that appealed only to conceptual truths and hence allowed no room for error. While (N)'s purity may lead us to believe that (B) is *a priori*, the application of (N) to particular contexts will soon convince us that whether truths are more useful to believe than falsehoods is an empirical issue, which is what one would expect from claims about utility.[12]

The above discussion has shown the implausibility of minimalistic accounts of one general claim that seems to involve truth, that truth is a goal of enquiry. But the point is a general one: minimalism can't handle *any* general claim that apparently involves truth. I'll show this by reflecting on the presentation by Anil Gupta.

---

[12] In my argument here, I've been inspired by Lynch, 2004. But Lynch makes an interesting argument that I don't employ. Lynch points out that Horwich is committed to infinitely many axiomatic normative claims of form (1). Hence, Horwich is committed to an extreme form of normative particularism when it comes to belief. Whether this is true or not depends on what we require of normative generality. Lynch says that particularism is implausible because it makes normative learning difficult, and denies that normative reasoning appeals to general principles. In this context, the point would be that we don't seem to have to rederive each instance of (1) on the basis of a brand new instance of (N), and that we should nevertheless be able to derive all instances of (1) from something, to account for why we know them. But it's not clear why the general principle can't be the skill of applying (N) to new instances of (1), and why normative learning can't have been the discovery of (N).

Gupta's argument begins with a characterization of deflationism as the conjunction of four theses. The first thesis, disquotationalism, is that the truth predicate is a device of disquotation. The second and third theses give the function of the truth predicate: we have a truth predicate so that we can express infinite conjunction or generalizations. The fourth thesis is essential. It states that the truth predicate serves its expressive functions *because* of its disquotational power.[13]

For Gupta, the issue turns on what it is for a sentence with the truth predicate to express an infinite conjunction or generalization. Consider these three sentences:

2) [Sky is blue and snow is white] and [Chicago is blue and snow is white] and...

3) ['Sky is blue' is true and snow is white] and ['Chicago is blue' is true and snow is white] and...

4) For all sentences x: [x is true and snow is white].[14]

This is a set of examples meant to stand in for more serious infinite conjunctions. Sentence (2) is the literal infinite conjunction. Sentence (3) is the same conjunction, but with semantic ascent by means of the truth predicate. Sentence (4) is the generalization abbreviated through the use of the truth predicate. So we are able to use (4) to express (2).

A more serious example might help us understand what's going on here. Consider something that we might want to say, like (4'):"If a claim has been confirmed using scientific methods, then it is likely to be true." According to minimalism, without the truth predicate, this sentence would go like (2'): "If *p* has been confirmed using scientific methods, then it is likely that *p*; if *q* has been confirmed using scientific methods, then it is likely that *q*; if *r*...." Thus the sentence is an infinite generalization or conjunction. It

[13] Gupta 1993, p. 287.

[14] *ibid*., p. 286. The eccentric numbering matches Gupta's, but I left out his totally uninteresting sentence (1). The numbering in the rest of this section disregards the numbering from before these sentences from Gupta.

can express that infinite content *because* of the disquotational power of the predicate 'is true.'

Let's consider the relationships between (2), (3), and (4). For minimalism to show that (4) expresses (2), there would have to be the appropriate sort of equivalence between each of (2) and (3) and (3) and (4). There are three grades of equivalence worth considering: mere material equivalence, necessary equivalence, and sameness of sense or meaning. If either of the equivalences is mere material, or even logical, equivalence, then it's hard to see how (4) could express (2).

Let me define an 'explanatory relata' of a sentence *p* as a sentence that either serves in a non-trivial way in an explanation of *p*, or for which there is an explanation in which *p* serves in a non-trivial way. Minimalism claims that truth is not an explanatory feature: it neither requires, nor provides, explanations. So any explanatory relata of (4) must equally well be an explanatory relata of (2), since (4) has as its function expressing (2). Assume that *q* is an explanatory relata of (4). We may agree that (2) and (3) are each true if and only if (4) is true, while denying that *q* is an explanatory relata of either (2) or (3). It's possible to explain one of two materially equivalent sentences without explaining the other. Likewise if the equivalences are taken to be logical or necessary. The only equivalence that's sufficient to give (4) and (2) exactly the same explanatory relata would be sameness of sense or meaning.[15]

But is it plausible that (4) and (2) have the same meaning? Not at all. To see this, we should check a more serious example. Recall:

> (4'): If a claim has been confirmed using scientific methods, then it is likely to be true.
>
> (2'): If *p* has been confirmed using scientific methods, then it is likely that *p*; if *q* has been confirmed using scientific methods, then it is likely that *q*; if *r*....

[15] *ibid*., p. 289.

It's much easier to explain (4') than it is to explain (2'). To explain (4'), we could use inductive or abductive reasoning about things with the property of truth to show that being confirmed using scientific methods tracks well with truth. But to explain (2'), we would have to explain each conjunct. Since we can explain (4') without explaining (2'), they don't have the same meaning, and (4') doesn't express (2'). We need the property of truth to figure, in a substantive way, in (4') for it to be explainable. Hence truth is substantive, not minimal.

Horwich's reply is very weak:

> Anil Gupta rightly notes that the instances of the generalizations that we use the concept of truth to formulate will not say exactly the same thing as what we wished to generalize unless corresponding instances of 'The statement *that p* is true' and '*p*' express the very same proposition — which... is not very plausible. But this point does not undermine the minimalist story about the function of truth; for... that function requires merely that the generalizations permit us to *derive* the statements to be generalized — which requires merely that the truth schemata provide material equivalences. (Horwich 1990/1998, p. 124)

Gupta's argument is that, if minimalism were true, then anything that explained (4) would have to explain (2), since minimalism claims that (4) is just (2) abbreviated. But it is possible to explain (4) without explaining (2). Hence, (4) is not just (2) abbreviated. Horwich's reply appears to be that the derivability of all of the conjuncts of (2) from the generalization (4) is sufficient for (4) to express (2). One gets the feel that Horwich is lagging behind in the conversation, or didn't follow Gupta's argument all the way through, because the possibility that (4) and (2) only need to be materially equivalent for the one to express the other is specifically dealt with in Gupta's argument. It's not clear what Horwich would say were he to respond to Gupta's argument in a more full-blooded way.

Horwich, however, denies that truth has any explanatory power. With respect to generalizations like (4'), he claims that, "Given the function of our concept of truth, we can see that these generalizations are not focused on truth, not really about truth. Rather they belong to that class of special schematic generalizations that rely on the equivalence

schema for their formulation." (*ibid.* p. 141) (4'), contrary to appearances, does not relate truth to confirmation by science. It's just another way of saying (2'). But, as we've just seen, it is patently *not* a way of saying (2'). The only way that we could be in a position to explain (2') would be to have the list of conjuncts in hand. But we could explain (4') on the basis of inductive or abductive reasoning about the connection between truth and science.

The point to be made might be even clearer if we set aside explanation and focus on justification. Under what circumstances might we be justified in claiming (2') and (4')? If we had its infinitely many conjuncts in hand, then we would be justified in claiming (2'). But, setting aside that that circumstance will never obtain, even if it did, we would not be justified in claiming (4'). Having a list of infinitely many propositions, asserting that, if it's confirmed by science, then it (is true), will not justify the universal generalization of those propositions. To have the generalization, we would need to have in hand the negative existential claim that there are no more propositions. So there is a circumstance that would justify (2') but not (4').

Let's try the other direction. If we had suitably many instances of (4'), and we thought that they supported the generalization, and that such a generalization would support counterfactuals, then we would be justified in claiming (4') (and hence (2')). But that's because we would be treating (4') as the statement of a law, not a mere universal generalization. If we were to set aside this feature of (4'), that is, set aside that it makes important reference to truth, then we would never be in a position to claim (4'). Yet not only *could* we be in a position to claim (4'), it seems that we *are* in a position to claim (4'). Hence truth is key to the justification of (4'), so truth must not be minimal.

1.1.5 Horwich's Use Theory

Why address the theory of meaning as use at all, much less in the section on truth, rather than on meaning? The doctrine that meaning is use makes a difference to our assessment of Horwich's minimalism. Minimalism thins out truth so much that the truth-

conditional account of meaning would be implausible were minimalism true, as Dummett explains:

> ...if we accept the redundancy theory of 'true' and 'false'.... we must abandon the idea which we naturally have that the notions of truth and falsity play any essential role in any account either of the meanings of statements in general or of the meaning of any particular statement. (Dummett 1959, p. 7)

The danger to my project is obvious: I intend to turn general arguments against scepticism and relativism that rely on the truth-conditional account of meaning into support for moral realism. Minimalism, by undercutting the account of meaning, would undercut my arguments against scepticism and relativism, and thence indirectly invalidate my project. But, on the other hand, since the minimalist undermines an attractive theory of meaning, he must offer his own in its place. But what if no such theory — one designed to be a natural fit with minimalism — were plausible? Then the minimalist would be unable to place his theory in the context of an overall philosophy of language, which is certainly a deficiency. This attack is, to be sure, *ad hominem*, but still, if successful, it points out a real problem with minimalism. So in this section I present Horwich's use theory. The presentation of minimalism in 1.1.2 was critical but did not attempt to close the case. This section will be different. I think that Horwich's use theory is susceptible to very quick refutation, and I provide it in this section.

Horwich's theory is not, precisely, a use theory of meaning, for Horwich does not claim that uses are meanings: "*Meanings are concepts*" (Horwich 1998, p. 44). Rather, he claims that what it is to have a given meaning is to be used in a certain way. The meaning itself, a concept or property, is independent.

Basic to Horwich's approach is his view of the relation between the logical forms of meaning claims (claims of the form, "x means F") and their reductions. As Horwich notes, a constraint he calls Strong Relationality is "satisfied by just about every theory [of meaning] that has ever been seriously entertained." (Horwich 1998, p. 23) The Strong Relationality Constraint is a constraint on reductive theories of meaning that they

preserve the logical form of the claims they reduce, which is obviously relational: to say that x means F is to relate x to F. Horwich rejects the Constraint:

> Notice... that this is a very substantive commitment, and that it stands in need of justification. For why should it not be, on the contrary, that although there is *perhaps* some relation to dogs that constitutes the property of meaning [the concept] DOG, this is completely different from the relation to tables that constitutes the property of meaning [the concept] TABLE? (ibid, p. 24)

Horwich's goal in (as distinct from his argument for) rejecting the Constraint is this. He wants to reduce "x means F" to facts about the use of 'x.' But 'x's being used in a certain way is not, apparently, a relational property; in any event, it does not seem to be a relation to a meaning. So if any reductive analysis of "x means F" had to retain the relational logical form of the original, then the use theory couldn't even get off the ground.

Horwich doesn't have an argument against the Constraint, but he argues that it requires justification that it has never received. There's nothing in the nature of reductive analyses, he points out, that requires them to preserve logical form. Let '*u(x),*' some underlying property, be a purported reduction of '*s(x),*' some superficial property. Then '*u(x)*' constitutes or reduces '*s(x)*' just in case:

(1) '*u(x)*' and '*s(x)*' apply to the same things, and

(2) facts about '*s(x)*' are explained by (1). (*ibid*, p. 25)

But nothing in these two conditions on reduction or constitution demands that superficial relations be constituted by underlying relations; perhaps what underlies the relation is a monadic property. I think that Horwich is right to say that the Strong Relationality Constraint is not a necessary constraint. However, there is at least one consideration in its favor.

To offer a reduction that differs in logical form from that which it reduces is to offer a substantially revisionist reduction. An underlying property *is* the property being underlain; being H_2O *is* being water. What makes one property "underlie" the other isn't

an ontological distinction, but an epistemic one. One property underlies another just in case it is that property, but understood in the context of a theory that reveals a great deal about the nature of that property. Being water doesn't explain much about water, since wateriness is not embedded in a powerful theory. Being H_2O, however, is embedded in a powerful chemical theory that explains a great deal about water. So replacing a superficial property with its underlying property is not to replace one thing with another, but to "replace" one thing with itself but in a more informative way.

Finding underlying properties, then, is a scientific endeavor. Science should be conservative and not reject established ideas without good reason. Now, suppose that the logical form of being water has feature A, while the logical form of being H_2O has feature B, and nothing with either feature has the other. Since being H_2O just *is* being water, to replace being water with an underlying property with a different logical form is to violate conservatism. Often, scientists must do this, but never without fairly good reason.

Likewise in the context of meaning. The fact that the meaning predicate is relational suggests that we ought to look for a relational underlying property. It may be that we won't find one, but the Constraint is a good heuristic device. So while Horwich is right to point out that we *may* set aside the Constraint, he is wrong not to pay the Constraint its due as a principle of scientific conservatism.

Here's a less abstract way to see why conservatism in the form of the Constraint ought to be respected if at all possible. Consider two meaning claims, "'a' means F" and "'b' means G." We would expect that these two claims share some content, the relation of meaning; this is suggested by the appearance in both claims of the 6-character string 'means_'. But for Horwich, "means F" is a monadic predicate. Thus it will be only an accident of spelling that it seems very much like the distinct predicate, "means G." But since they have nothing in common, there is no longer any reason to believe that an analysis of one of them will at all resemble an analysis of the other; for instance, that one

of them can be analyzed with reference to its use won't even begin to hint that the other one can be analyzed with reference to *its* use. Since no two meaning claims have anything at all in common, Horwich, to persuade us that his theory is correct, will have to give an independent reductive analysis of each and every meaning claim that we could make. Only by obeying the Constraint can we be in a position to offer a theory of meaning, as distinct from a theory of "means F" and a different theory of "means G," and so forth, which, for no apparent reason, happen to look a lot alike.

Horwich's reducing meaning_F to use but at the same time positing meanings is puzzling. What is the sense in positing meanings? Intuitively, one of the advantages of the use theory of meaning is that it does away with mysterious relata of words; use looks more amenable to physicalistic reduction than do intensionalistic meaning posits. Why give up this advantage? What does introducing meanings do for Horwich? Davidson remarked, topically: "Paradoxically, the one thing meanings do not seem to do is oil the wheels of a theory of meaning.... My objection to meanings in the theory of meaning is not that they are abstract or that their identity conditions are obscure, but that they have no demonstrated use." (Davidson 1967, pp. 20-1) Horwich does make propositional attitudes consist of concepts (meanings) stuck together, but this claim isn't very informative about attitudes: it just says that the meaning of 'Fa' is the meaning of 'a,' plus the meaning of 'F.' Perhaps it's enlightening to be told that my mental contents are independent entities that exist in a Platonic realm to which other people have access, but if that's what's interesting, Horwich spends little time on what's interesting.

The monadic reduction of meaning_F is troubled, and all the detritus of relational theories of meaning appears still to be in place. But why is the reduction of meaning_F to use a good one? Horwich explains:

> *The overall use of each word stems from its possession of a basic acceptance property*. For each word there is a small set of simple properties which... explain total linguistic behavior with respect to that word. These explanatorily basic properties fall into various kinds... where each such kind is defined by the distinctive form of its members and by the range of phenomena they are needed to

> account for. The present theory is focused on the *semantic* feature of a word. The distinctive form of that feature is that it designates the circumstances in which certain specified sentences containing the word are accepted; and the primary explanatory role of a word's acceptance property is to account for the acceptance of other sentences containing the word.
>
> ... for each word *w*, there is a regularity of the form
>
> All uses of *w* stem from its possession of acceptance property *A(x)*,
>
> where *A(x)* gives the circumstances in which certain specified sentences containing *w* are accepted. (Horwich 1998, pp. 44-5)

Horwichian semantics won't be like Tarskian semantics. For Tarski, semantics is a relational discipline; words are related to objects. But for Horwich, semantics seems to be (more or less by definition) a non-relational discipline. It specifies certain paradigm sentences in which a word is used, acceptance of which defines what it is for a word to have its meaning. That a speaker accepts those sentences (plus various worldly phenomena) commits her to the acceptance of most other accepted sentences that involve the word, and thus explains why she is inclined to utter or at least accept them in the appropriate contexts.

Key, from my point of view, are the relations between the use theory and truth. Davidson tries to soften the opposition between the use and truth-conditional theories of meaning:

> What is clear is that someone who knows under what conditions a sentence would be true understands that sentence, and if the sentence has a truth value, then someone who does not know under what conditions it would be true doesn't understand it. This simple claim doesn't rule out an account of meaning which holds that sentences mean what they do because of how they are used; it may be that they are used as they are because of their truth conditions, and they have the truth conditions they do because of how they are used. (Davidson 1997, p. 13)

On this possible account, use and truth-conditions are mutually determining, and use determines meaning. Such an account wouldn't be functionally very different from one that has it that use and truth-conditions are mutually determining, and truth-conditions determine meaning. Horwich, however, places the disagreement:

> ...even if Davidson were to accept a use theory of *truth conditions*, this approach would conflict on the question of explanatory order with the use theory of *meaning*. For on the latter view a sentence's truth condition is a *consequence* of its meaning, not *constitutive* of it. (Horwich 1998, p. 72)

This puts the difference between the use theory and the truth conditional theory very sharply. If Horwich can show that knowledge of truth conditions flows from knowledge of meaning (otherwise than by being the same thing), then all's well. Horwich says that, on the use theory, we discover a truth-condition through a 3-stage procedure:

> First, we know the meaning of "snow is white" by knowing its mode of construction and the uses of its component words. Second, we know the meaning of "*true*" by accepting instances of 'The proposition that *p* is true iff *p*' and accepting '(*u*)[*u* is *true* iff (∃*x*)(*u* expresses *x* & *x* is true)]', and then inferring instances of the disquotation schema, '"*p*" is *true* iff *p*' — including '"snow is white" is *true* if and only if snow is white.' And third, in so far as we understand all the constituents of that biconditional, we can be said to know *that "snow is white" is true if and only if snow is white*. Thus our knowledge of the truth conditions of "snow is white" *derives from* our knowledge of its meaning. (*ibid*)

The third phase isn't very interesting; it's the phase at which we point out what we've already accomplished. The second phase, note, involves *inference*. The derivation appears to be logical; the appropriate instance of the disquotation schema is derived from other items of knowledge through an argument. I don't believe that this argument works, so I'll look at it very closely.

How shall we represent the knowledge of the meaning of whatever sentence we're talking about; a, for instance? This will turn out to be crucial and I'll have to discuss it at length.

The three phrases, "the proposition that *p*," "*p*", and *p* all look much more similar than they should. The first is the same of an abstract entity, the second is the name of a sentence, and the third is a sentence. To make things clearer, I introduce 'a' to name a sentence, 'b' to name a proposition, and 'A', to be a sentence. 'A' is the sentence that a names, and b is the proposition that it expresses.

Now, let's look at the most formally expressed of the premises: (*u*)[*u* is *true* iff (∃*x*)(*u* expresses *x* & *x* is true)]. Between two eventual names that will replace '*u*' and '*x*,' there is the relation of expression. We may take it that whatever will replace '*x*' is intended to name a proposition, while whatever will replace '*u*' is intended to name something that will express a proposition. But now note that nothing is to be true unless it expresses a proposition that is true. Hence, for a proposition to be true, it must express a proposition that is true. We need to make clear that '*u*' is to be replaced by a sentence.

Our grasp of truth, phase 2, seems to consist in our grasping two claims; an instance of "The proposition that *p* is true iff *p*," and some suitably modified version of the very formal-looking claim. Let the first of these be premise (1) Tb ↔ A. b, recall, is the proposition expressed by 'A.' That proposition is to be true just in case A. Let the second be premise (2) ∀*u*{S*u* → [T*u* ↔∃*x*(*u*E*x* & T*x*)]}. For any sentence, that sentence is true just in case it expresses something that is true. We'll take for granted that only propositions are ever expressed. Finally, what we're trying to discover here is the truth conditions of a sentence like "Snow is white." So our conclusion, hopefully, will be that a (i.e., 'A') is true iff A: (C) Ta ↔ A.

It's readily apparent that our argument cannot get off the ground, but that's probably because we haven't yet encoded our grasp of the meaning of a. That should be premise 3. How shall we express this knowledge? On a first glance, the argument will probably require some claims about what *expresses* what, since the second premise includes the relational predicate, 'expresses.' But what is expression? Horwich never exactly says, but I think his meaning comes out clearly here: "*Two words express the same concept in virtue of having the same basic acceptance property*." (*ibid*, p. 46) Concepts are meanings, and meanings are determined by acceptance properties. Yet here words are said to express concepts. I guess that 'express' and 'mean' express the same concept.

Premise 3 is supposed to encode the speaker's grasp of the meaning of a. But the meaning of a is b. Meanings are what are expressed, so we may introduce our first pass at premise (3) aEb. We may also state the trivial: (4) Sa, a is a sentence.

The strategy is plain; we're trying to prove a biconditional claim, so we assume each side of the biconditional and prove the other side in a conditional proof.

Based on our premises, we can prove that A → Ta:

(1) Tb ↔ A	
(2) ∀u{Su → [Tu ↔ ∃x(uEx & Tx)]}	
(3) aEb	
(4) Sa	
(5) Sa → [Ta ↔∃x(aEx & Tx)]	∀, 2
(6) Ta ↔∃x(aEx & Tx)	→, 4, 5
(7) A	Assume for →
(8) Tb	↔, 7, 1
(9) aEb & Tb	&, 3, 8
(10) ∃x(aEx & Tx)	∃, 9
(11) Ta	↔, 10, 6
(12) A → Ta	→, 7, 11

But the reverse cannot be made to work:

(13) Ta	Assume for →
(14) ∃x(aEx & Tx)	↔, 6, 13
(15) aEc & Tc	∃, 14
(16) Tc	&, 15
(17) A	↔, 16, 1

To get A, we must have Tb, and to get Tb, we must instantiate the existential on line 14 to b. But we can't, due to the constraint that a constant introduced on the basis of an existential be new to the proof. One might think to try a *reductio*:

(15) ~A	Assume for *reductio*
(16) ~Tb	↔, 1, 15
(17) aEb & Tb	&, 3, 16
(18) ∃x(aEx & ~Tx)	E, 17
(19) ~Ta	contradiction 14, 18
(20) A	contradiction, 13, 19

But (18) doesn't contradict (14); no contradiction seems to be in the offing. If a is true (i.e., 'A' is true) and 'A' is false, then there's something that a expresses that is false, but that doesn't contradict the claim that there's something that a expresses that is true. We

need to make it clear that the proposition that a expresses is unique; a expresses only one proposition: we need to change premise (3). Currently, (3) only says that a expresses b, leaving open the possibility that there might be many things that a expresses, some true but some false. We must restrict a's expressive power, so that it may only express the one proposition: $\forall x(aEx \rightarrow x=b)$. This won't do either. By instantiating 'x,' we find only that, if a expresses anything, it's b. But perhaps a doesn't express anything. We need to assert that a does express exactly one thing: (3R) $\exists x[aEx \ \& \ \forall y(aEy \leftrightarrow x=y)]$. But this will force a change to premise (1). We will wish to instantiate 'x' in (3R), but as long as (1) has mention of b, we won't be able to instantiate 'x' to b, which is what we need to do. But (1) must, in some way, identify the proposition expressed by a. If we use any name, the argument will be blocked. How can we identify the proposition otherwise than by naming it? We may introduce a definite description. All we know about b is that it is the proposition expressed by a, so let us identify it as: $\iota x(aEx)$, the thing such that a expresses it. Premise (1'), then, will be: $T\iota x(aEx) \leftrightarrow A$: the thing expressed by a is true iff A.

But now that we have this definite description in place, we may notice that (3R) is a sentence with a definite description, expanded out according to Russell:

(3R) $\exists x[aEx \ \& \ \forall y(aEy \leftrightarrow x=y)]$

is what's meant by:

(3') $aE[\iota x(aEx)]$

that is, a expresses the thing expressed by a. Since we are working in an extensional context, it doesn't strictly matter whether we expand the descriptions or not; I find it easier to leave the descriptions in iota-notation and introduce a new proof rule ι:

$\phi[\iota x(\phi x)]$
$\phi\alpha$
$\therefore \alpha=\iota x(\phi x)$

The idea should be pretty plain. The thing of which ϕ is true is identified twice, once by name but once by definite description. But the description is: the thing of which ϕ is true.

It's plain, then, that anything of which ϕ is true must be that thing. With the name 'b' removed and the definite description in place, and proof rule ι, the argument works:

(1') Tιx(aEx) ↔ A	
(2) ∀u{Su → [Tu ↔ ∃x(uEx & Tx)]}	
(3') aEιx(aEx)	
(4) Sa	
(5) Sa → [Ta ↔ ∃x(aEx & Tx)]	∀, 2
(6) Ta ↔ ∃x(aEx & Tx)	→, 4, 5
(7) A	assume for →
(8) Tιx(aEx)	↔, 1', 7
(9) aEιx(aEx) & Tιx(aEx)	&, 3', 8
(10) ∃x(aEx & Tx)	∃, 9
(11) Ta	↔, 6, 10
(12) A → Ta	→, 16, 20
(13) Ta	assume for →
(14) ∃x(aEx & Tx)	↔, 6, 13
(15) aEb & Tb	∃, 14
(16) aEb	&, 15
(17) b=ιx(aEx)	ι, 3', 16
(18) Tb	&, 15
(19) Tιx(aEx)	=, 17, 18
(20) A	↔, 1', 19
(21) Ta → A	→, 13, 20
(22) Ta ↔ A	↔, 12, 21

This argument is valid. In the first pass at the argument, we were able to prove A → Ta; we were able to prove that here without appeal to the new proof rule ι. That rule was important in deriving the other direction, Ta → A, as one would expect.

But how interesting is this argument from Horwich's point of view? Recall that the point of this project was to show how one grasps the truth conditions of sentences. But let's consider the substance of the premises. (4) is unexceptionable, and (2) seems reasonable. But (3') is trivial (as the silliness of the ι rule shows). It says only that a sentence expresses the proposition that it expresses. That is not a substantial claim; it does not actually identify the proposition. Our identification of that proposition has been "linguistic through the phrase." We were supposed to be basing our knowledge of sentences' truth-conditions on substantial knowledge of meanings, but no such substantial

knowledge is present. Likewise, (1') is trivial: it says only that the proposition expressed by a sentence 'A' is true iff A. The situation here recalls the Russellian argument against propositions from section 1.1.3.

But perhaps my repair of the argument has been improper. Instead of replacing b with a definite description and introducing proof rule ι, I should have introduced a new premise that states explicitly that a expresses only one proposition: (5) ∃!x(aEx). This will come with a proof rule. Here's rule !:

$\phi\alpha$
$\phi\beta$
$\exists!x(\phi x)$
$\therefore \alpha=\beta$

The sense should be clear. We count two things of which ϕ are true, but we know that there is only one such thing; hence, the things we counted must actually be identical.

Returning to our original premises that included the name 'b' of the proposition expressed by a, but introducing a new premise (5) stating that there is exactly one such proposition, we can again make the argument work:

(1) Tb ↔ A	
(2) ∀u{Su → [Tu ↔ ∃x(uEx & Tx)]}	
(3) aEb	
(4) Sa	
(5) ∃!x(aEx)	
(6) Sa → [Ta ↔∃x(aEx & Tx)]	∀, 2
(7) Ta ↔∃x(aEx & Tx)	→, 4, 6
(8) A	Assume for →
(9) Tb	↔, 8, 1
(10) aEb & Tb	&, 3, 9
(11) ∃x(aEx & Tx)	∃, 10
(12) Ta	↔, 11, 7
(13) A → Ta	→, 7, 12
(14) Ta	Assume for →
(15) ∃x(aEx & Tx)	↔, 7, 14
(16) aEc & Tc	∃, 15
(17) aEc	&, 16
(18) b=c	!, 3, 5, 17
(19) Tc	&, 16
(20) Tb	=, 18, 19
(21) A	↔, 1, 20

(22) Ta → A →, 13, 21
(23) Ta ↔ A ↔, 12, 22

As with the previous argument, it is possible to show that A → Ta without appeal to anything new. But we must appeal to premise (5) and the new proof rule ! to show that Ta → A. For the truth of a sentence to imply the sentence, we must assume that the sentence expresses only one proposition. But it's not plain that we're in a position to accept that. Horwich explains:

> ...there is the question of which of the many such properties of a given word is to be the one associated with its meaning. The right answer, I shall argue, is that we want the property that is *explanatorily basic*.... if we think of this as a generalization regarding the use of the word, then the one we choose is the one that provides the simplest account of all the word's individual uses. The fact that there may well be no such thing — but only a range of equally good choices — is what constitutes the *indeterminacy of meaning*. (*ibid*. p. 41)

For the argument to go through, each sentence with truth conditions must be associated with exactly one proposition, but as Horwich admits, that is not the case. A sentence is associated with a range of propositions, and there's no fact of the matter which one it expresses.

Perhaps this suggests another reformulation of the argument. Now, the problem appears to be with premise (2): ∀u{Su → [Tu ↔ ∃x(uEx & Tx)]}. So far, we have required that there be exactly one proposition expressed by a. But maybe we should rewrite (2) so that it asserts that a expresses at least one proposition, and that a is true iff *every* proposition a expresses is true: (2') ∀u{Su → [Tu ↔ ∃x(uEx & ∀y(uEy → Ty))]}. Then we could return to the original premises (1) and (3)-(4). We can prove Ta → A, which is the opposite of the direction were able to show on the original, unreformed, premises.

(1) Tb ↔ A
(2') ∀u{Su → [Tu ↔ ∃x(uEx & ∀y(uEy → Ty))]}
(3) aEb
(4) Sa
(5) Sa → {Ta ↔ ∃x[aEx & ∀y(aEy → Ty)]}
(6) Ta ↔ ∃x[aEx & ∀y(aEy → Ty)]

(7) Ta — assume for →
(8) ∃x[aEx & ∀y(aEy → Ty)] — →, 6, 7
(9) aEc & ∀y(aEy → Ty) — E, 8
(10) ∀y(aEy → Ty) — &, 9
(11) aEb → Tb — ∃, 10
(12) Tb — →, 3, 11
(13) A — ↔, 1, 12
(14) Ta → A — →, 7, 13

However, we will not be able to prove A → Ta, which was provable on every other version of the argument:

(15) A — assume for →
(16) Tb — →, 1, 15
    (17) ~Ta — assume for *reductio*
    (18) ~∃x[aEx & ∀y(aEy → Ty)] — ↔, 6, 17
    (19) ∀x~[aEx & ∀y(aEy → Ty)] — ~∃, 18
    (20) ~[aEb & ∀y(aEy → Ty)] — ∀, 19
    (21) ~aEb ∨ ~∀y(aEy → Ty) — ~&, 20
    (22) ~∀y(aEy → Ty) — ∨, 3, 21
    (23) ∃y~(aEy → Ty) — ~∀, 22
    (24) ∃y(aEy & ~Ty) — ~→, 23
    (25) aEb & Tb — &, 3, 16
    (26) ∃y(aEy & Ty) — ∃, 25
    (27) Ta — contradiction, 24, 26

But 24 and 26 don't contradict and I don't see any other contradiction on the horizon. There seems to be no reason not to accept A while rejecting Ta. Horwich's argument won't go through unless we assume that each sentence expresses exactly one proposition. Thus, Horwich's semantics requires too much determinacy; it can't accept even the modest indeterminacy that he accepts. The situation here recalls the Quinean argument against propositions from section 1.1.3.

So far, I've been very accepting of Horwich's premises. But there is a major contradiction lurking within them. Recall that Horwich presented fairly clear versions of premises (1) and (2). But I had to do some interpretive work to arrive at premise 3: aEb. Horwich's sentence that I tried to encode with premise (3) was: "First, we know the meaning of "snow is white" by knowing its mode of construction and the uses of its

component words." What you know when you know the meaning of a, is that a expresses b, I have been claiming.

But this interpretation conflicts with Horwich's opposition to the Strong Relationality Constraint. We should not, according to Horwich, represent a sentence's having a meaning by relating to its meaning. Rather, having that meaning is a monadic property. So (3) should have been: Fa, given some suitable F. So far, this isn't troubling, but what are we to make of (2): $\forall u\{Su \rightarrow [Tu \leftrightarrow \exists x(uEx \ \& \ Tx)]\}$? I made up premise (3) so that it would coordinate with the 'uEx' of (2). It's fairly easy to rewrite (3), but what can be done with (2)? Anything that does away with the relation between 'u' and 'x,' and hence satisfies the denial of Strong Relationality, will also unhook the sentence from truth. To make the argument work, we need the truth of the sentence to connect with the truth of a proposition, and to make that connection, we need to connect the sentence with the proposition. But that is the thing that we may not do if we deny Strong Relationality. Horwich accepts all of this: "...the use theory does indeed violate the commonly assumed requirement that there be explanations of the links between given meaning-constituting properties and given meanings. But this requirement is misconceived; so our violating of it is not objectionable." (*ibid*, p. 66) This claim requires consideration.

Let F(x) be the meaning-constituting property of a, and let b be the meaning of a. What Horwich is now telling us is that Fa has no "explanatory links" to b, or, presumably, to aEb: "...as we saw..., the expectation that one will be able to read off (and hence to explain) which particular meaning is engendered by a given meaning-constituting property is misguided." (*ibid*, p. 220) According to Horwich, knowing Fa will not help you learn that aEb. But recall that what it is for one property to underlie another is for them to be co-extensional and for the one to explain facts about the other. If having a given meaning-constituting property doesn't even determine which meaning a word has, how can it possibly account for any feature of that meaning?

In particular, consider the result of deepening the relatively superficial premises (2) and (3). (3) turns into (3"): Fa. But (2), shorn of reference to the superficial relation of expression, is inconceivable. Strangely, if we try to appeal to the meaning-constituting properties, we will no longer be in a position to grasp sentences' truth-conditions; to grasp sentences' truth-conditions, we have to remain at the superficial level and treat meaning as relational. As soon as we actually commit to the signature thesis of Horwich's use theory, we will no longer be able to grasp sentences' truth-conditions. Surely this theory fails as badly as it is possible for a theory of meaning to fail.

The point of the argument I've been analyzing was to show how we can learn sentences' truth-conditions on the basis of a grasp of their meanings. The point of showing that was to demonstrate the contrast between Davidson's truth-conditional semantics and Horwich's use-theoretic semantics. For Horwich, the difference between the theories is in order of explanation: one accounts for uses on the basis of truth-conditions, while the other accounts for truth-conditions on the basis of uses. We've seen how very badly things went when Horwich tried to account for truth-conditions on the basis of uses. But Horwich ties his approach to Davidson's in a different way: "once its precise content is elaborated, Davidson's Principle of Charity arguably boils down to the use theory of meaning." (*ibid*, p. 72) The Principle of Charity is dear to my heart, and since the use theory of meaning looks like a disaster, I hope that the one doesn't boil down to the other. What does Horwich have in mind? Recall which regularities of use constitute meaning:

> ... for each word *w*, there is a regularity of the form
>
> All uses of *w* stem from its possession of acceptance property *A(x)*,
>
> where *A(x)* gives the circumstances in which certain specified sentences containing *w* are accepted. (*ibid*, pp. 45)

It's plain why Horwich sees the similarity.

For Davidson, speakers exhibit the attitude of holding sentences true. In interpretation, we must optimize agreement between the sentences we hold true, and the sentences the speakers hold true. If they hold one of their sentences true under the same circumstances under which we would hold one of our sentences true, then the two sentences mean the same.

For Horwich, speakers exhibit the attitude of acceptance. In interpretation, we must check for identity of acceptance conditions. If they accept one of their sentences under the same circumstances under which we accept one of ours, then the two sentences mean the same.

The two accounts do look to be more or less the same. But while Horwich is trying to reduce Davidson's account to his own (Davidson's account 'boils down' to the use-theoretic account), the natural thing for Davidson to do is reduce Horwich's to his own. Recall that Davidson's third objection was that Horwich claims that meanings are constituted by acceptance conditions, but to accept something is to accept it as true. Hence, the concept of truth is a prerequisite for grasp of Horwichian meanings. Horwich replies:

> Granted, *accepting* a sentence goes hand in hand with accepting its *truth*. But, equally well, *supposing* something goes hand in hand with supposing its *truth*, *doubting* something goes hand in hand with doubting its *truth*, and so on.... Consequently, its relationship to truth is not what distinguishes acceptance from other attitudes... and does not help to constitute its nature. Thus the relevant concept of acceptance does not presuppose the notion of truth. (Horwich 1998, p. 95)

Horwich says that the fact that acceptance is acceptance as true doesn't imply that acceptance is based on truth, since that other attitudes are doubting to be true or hoping to be true doesn't imply that doubting or hoping are based on truth. But of course on the truth-conditional account, they *are* based on truth: you can't hope that p without knowing p's truth-conditions.

Horwich presents minimalism and the use theory as structurally similar to Davidson's theories of truth and meaning, which they are, but as differing in explanatory order. The core of the view, then, is the derivation of a sentence's truth-conditions from a grasp of the meaning of the sentence, plus the meaning of the truth predicate. That derivation totally failed. At every turn, it displayed one of the failures of Horwich's theory. First, that it commits itself to propositions but gives no criteria by which to identify them. Second, that it commits itself to overdeterminate meanings while admitting indeterminacy. Third, that its meaning-constituting properties are devoid of explanatory power. All of this shows that the correct order of explanation runs from truth-conditions to uses, not the other way around. Truth is a substantial property with explanatory power.

1.2 TRADITIONAL CORRESPONDENCE THEORIES

1.2.1 Universals and the Slingshot

A correspondence theory of truth is one that defines truth as correspondence to a fact, so the correspondence theorist must give theories of correspondence and facts. I don't worry about correspondence. I use the term ‘fact’ in a way that’s more respectful of the history of philosophy than it is of ordinary usage: I take facts to be controversial entities; anything that I'm willing to call a fact will have to bear some sort of similarity to the sort of things talked about by Russell, early Wittgenstein, and other classical correspondence theorists. The similarity has to be in the content of the idea of a fact, not in the purpose for which the idea is invoked. Not just anything that makes sentences true can count as a fact.

We can divide theories of facts into two kinds, those that make universals (properties, relations) constituents of facts, and those that don't. Neither kind of theory will be acceptable. In this section, I employ the formal argument known as the Slingshot to show that, if universals are not constituents of facts, there will be far fewer facts than the correspondence theorist needs for her view to be remotely plausible. In the next two

sections, I consider Davidson's arguments concerning the unity of the proposition — which amount to the Bradley Regress — and argue that theories that satisfy the Slingshot will fail to provide unity for their facts. First, I present a naïve correspondence theory that will immediately fail. Second, I present a sophisticated recent view offered by Hochberg and derived from Russell; this theory will also fail, though it's somewhat more difficult to see why. The point of looking at the naïve theory is to see the pattern that will recur in the analysis of the sophisticated view. The two theories are variations on a theme.

The Slingshot is a classical argument, but I take the recent presentation by Stephen Neale to be the most authoritative treatment to date. The Slingshot will take four dummy premises:, Fa, Gb, a≠b, and ❤Fa, where '❤' is some one-place sentence connective. We then test two assumptions: that sentences within the context of '❤' are +ι-subs, and +ι-conv. To be +ι-subs and +ι-conv is for the proof rules of ι-subs and ι-conv to be validly employed within sentences within the ❤ connective.

Since I'm using the Slingshot to study facts, our ❤ connective works like this:

❤Fa is true if and only if 'Fa' is made true by the fact that p

❤ is defined, then, with reference to the fact that p. The Slingshot itself, as a formal argument, can take any one-place sentential connective, but I'm not concerned with the Slingshot as a formal argument. I want to apply it to the particular case of facts.

The Slingshot tests whether we may apply ι-conv and ι-subs within the ❤ context, so we must state these proof rules. ι-subs is these three rules:

$\iota x(\phi x) = \iota x(\psi x)$	$\iota x(\phi x) = \alpha$	$\iota x(\phi x) = \alpha$
$\Sigma\iota x(\phi x)$	$\Sigma\iota x(\phi x)$	$\Sigma(\alpha)$
$\therefore \Sigma\iota x(\psi x)$	$\therefore \Sigma(\alpha)$	$\therefore \Sigma\iota x(\phi x)$

The intuitive idea of these three rules that allow the substitution of one definite description for another or for a name should be fairly clear. Wherever Σ is an extensional predicate and none of the premises is within the context of a non-extensional sentence

connective, these are valid, as Russell and Whitehead showed.[16] ι-conv is the following pair of rules:

ι-intr: $T[\Sigma(x/\alpha)]$	ι-elim: $T[\alpha=\iota x(x=\alpha \wedge \Sigma(x))]$
$\therefore T[\alpha=\iota x(x=\alpha \wedge \Sigma(x))]$	$\therefore T[\Sigma(x/\alpha)]$

Here, the intuitive idea may be less clear. $\Sigma(x)$ is to be some formula with at least one instance of x and in which Σ is an extensional predicate; $\Sigma(x/\alpha)$ is $\Sigma(x)$ with every instance of x replaced by α; and $T\phi$ is just some sentence in which ϕ appears; for instance, ϕ, $\phi \rightarrow \psi$, or $\heartsuit\phi$.[17] An example may make things clearer. Let $\Sigma(x)$ be: x is a bachelor. Let $\Sigma(x/\alpha)$ be: Xander is a bachelor. And let $T\phi$ be: It's very disheartening to see that ϕ, so $T[\Sigma(x/\alpha)]$ will be: It's very disheartening to see that Xander is a bachelor. It seems that, if Xander is a bachelor, then *the thing that is Xander and is a bachelor* is a bachelor. We may then replace Xander's name with a description: 'the thing that is Xander and is a bachelor', or 'ιx(x=Xander & x is a bachelor)'. But, if it's disheartening to see that Xander is a bachelor, then surely it's disheartening to see that the thing that is Xander and is a bachelor, is a bachelor, and contrariwise. But that is ι-intr and ι-elim. In general, if we identify a thing by name and say something about it, then surely we could identify the thing by describing it as the thing that is it and of which whatever we wanted to say about it is true; it seems that such identification is only more tedious than identification by name but will have the same semantic and cognitive value.

The Slingshot shows that if $\heartsuit$ is +ι-subs and +ι-conv, then it also allows the substitution of materially equivalent non-general sentences. Hence, if Fa $\leftrightarrow$ Gb and $\heartsuit$Fa, then $\heartsuit$Gb. But let's assume that Fa and Gb are, intuitively, not made true by the same fact. Then there is a problem, since one of them being made true by a given fact implies that the other is made true by the same fact; in short, that there's only one non-general fact. No plausible correspondence theorist is going to invoke only one non-general fact.

---

[16] See Neale 2001, p. 160
[17] *ibid*, p. 179

What I need to show is that, on any theory of facts that denies universals, ❤ will turn out to be +ι-subs and +ι-conv. Unfortunately, I have no knock-down argument, but I think that denying universals and also at least one of ι-subs and ι-conv is going to make for a fairly unnatural-seeming theory. First, I consider ι-subs. ι-subs is a rule dealing with definite descriptions; it tells us to treat them more or less as names. Consider this argument:

Fa
a=b
❤Fa
∴❤Fb

That seems pretty harmless, assuming that 'F' is extensional. We are told that a is F, and that a is b. Then we note that 'Fa' is made true by the fact that p. But surely, the fact that makes true Fa will also make true Fb, since they make the same attribution to the same thing, albeit in different words. So it seems right to accept that, if Fa is made true by the fact that p, then so is Fb.

But consider this variant:

Fa
a=ιx(Gx)
❤Fa
∴❤Fιx(Gx)

Recall that we have rejected universals. It's hard to see, then, what the difference is between the contribution of 'a' and the contribution of 'ιx(Gx)' to the constitution of the fact could be. They seem to each contribute the same object, a. And as the previous argument showed, if two linguistic items contribute the same object, we will allow them to be intersubstituted within the ❤ context. So it's hard to see how we can resist accepting ι-subs.

Let me try an example. Assume that we agree that the fact that Xander is a bachelor is made true by the fact that p. Now, we want to figure out which fact to check for in considering whether the male slayerette is a bachelor, while bearing in mind that Xander is the male slayerette. It seems that we should check for the very same fact; we'll

check a certain entity, Xander, the male slayerette, and see whether he's a bachelor. It's hard to find room for additional facts. Likewise if, instead of 'Xander' and 'the male slayerette,' we have 'the male slayerette' and 'Buffy's best natural friend,' since the latter two are the same, there will again be only one fact to make true both of, "The male slayerette is a bachelor" and "Buffy's best natural friend is a bachelor." Where would we look for additional facts?

A contrast might help. Assume that we accept universals. Then perhaps the descriptions contribute, not Xander, but one or more universals. For instance, instead of contributing Xander, 'the male slayerette' might contribute maleness and the property of being a slayerette. In that case, there could be a semantic difference between a name and a definite description satisfied by the referent of that name, or two co-satisfied definite descriptions. Denying universals seems to cut off that possibility, so it seems that denying universals does make '❤' +ι-subs.

How about ι-conv? Similar considerations apply. Consider this argument:

❤ Fa
∴❤ a = ιx(x=a & Fx)

'Fa' is made true by the fact that p. But where else could we look for the fact that makes it true that a is the thing that is a and of which 'F' is true? Could such a fact be distinct from the fact that p? Consider the example of Xander's bachelorhood. That Xander is a bachelor seems to be the very same fact as the fact that Xander is the thing that is Xander and a bachelor. For Xander's identity, and his bachelorhood, are both present in the former fact, and the latter, and nothing seems to have been lost or added in the transition. It's just hard to see what the denier of universals who accepts facts is going to say here; at a minimum, I think that it's up to such a theorist to offer a positive theory of the difference between the truth-making facts for the two sentences.

But if we agree that, on a theory of facts that denies universals, '❤' is +ι-subs and +ι-conv, then the Slingshot will show that '❤' allows for the substitution of materially

equivalent non-general sentences. The argument properly speaking has four parts. For each part, we assume that Fa ↔ Gb and ❤Fa. For the first two parts, we assume Fa; for the first part, we say a=b, whereas for the second, a≠b. For the second two parts, we assume ~Fa; for the third part, we say a=b, whereas for the fourth, a≠b. All of the possibilities have been covered. I spare actually presenting the argument; those curious can consult Neale 2001, pp. 183-6 for a formal presentation. The upshot is that {Fa ↔ Gb, ❤Fa, ❤+ι-subs, ❤+ι-conv} implies ❤Gb. But ❤Fa said that 'Fa' was made true, if true, by the fact that p, and now, knowing only that 'Gb' has the same truth value as 'Fa,' we can see that it is made true by the very same fact. Since 'Fa' and 'Gb' are arbitrary non-general sentences, we can see that there are at most two non-general facts: the one that makes true non-general sentences true, and the one that makes false non-general sentences false.

Throughout, I have discussed the Slingshot only in application to non-general sentences. While I don't know of any good reason why the Slingshot can't be applied to general sentences, I don't know how we would begin to apply the crucial ι-subs and ι-conv rules to sentences like ∀x(Fx ↔ Gx) ↔ ∃y(Hy & Jy) and ❤∀x(Fx ↔ Gx) to show that ❤∃y(Hy & Jy). There are no names to convert into definite descriptions.

What if general truths like these are Slingshot-proof? That would encourage the suggestion that universals are what's necessary to block the Slingshot. The problem with non-general sentences, we might see, is that they include names and hence introduce particularity. Particulars can be definitely described, and thus allow their names to be converted into definite descriptions and for those descriptions to be substituted for one another. But sentences that make no reference to particulars, but only to general entities, might be not be susceptible to Slingshot-style reductions.

1.2.2 The Naive Correspondence Theory and the Unity of the Proposition

What I will call the 'Naive Correspondence Theory' is your basic correspondence theory that makes facts consist of particulars and universals, and that also accepts

Russell's Theory of Descriptions. With universals and Russell's theory in hand, it's easy to deny ι-conv and ι-subs.

Consider ι-conv. This rule allows us to convert a name in a sentence into a definite description without changing the content of the fact that would correspond to the sentence. But with universals in hand, we can immediately see the difference between the contributions made by names like 'a' and definite descriptions like 'ιx(Gx).' The latter does not contribute a to the sentence. It contributes the universal G-ness. So the fact that would correspond to 'Fa' would consist in a and F-ness, while the fact that would correspond to 'Fιx(Gx)' would consist in F-ness and G-ness (all facts to be suitably quantified). These facts are obviously different because they have different constituents, so if we accept universals and Russell's theory, it will be very natural to make ❤ -ι-conv.

Consider ι-subs. This rule allows us to substitute co-satisfied definite descriptions without changing the content of a sentence containing either description. But if the content of a description is a set of universals, and not a particular, then different descriptions will have different content even if they're co-satisfied. Assume that ιx(Gx)=ιx(Hx), and that Fιx(Gx). The latter fact consists in F-ness and G-ness, suitably quantified, and has nothing to do with H-ness, even though the only thing with G-ness is the only thing with H-ness. So substituting the descriptions within the latter sentence will produce a sentence that, though still true, would correspond to a different fact. So it's natural that ❤ will be -ι-subs.

Since Russell's theory is plausible on other grounds, the naive correspondence theory only really goes out on a limb in accepting universals. But the attractions of the correspondence theory are substantial; they're probably worth getting over allergies to universals. Unfortunately, though, appealing to universals in semantics leads to a different problem, the Bradley Regress.[18]

[18] This is not to say that universals, themselves, are a problem. It could be that universals exist; maybe they should appear in our best ontology. But they're useless for semantics, so the motivation to accept them should be non-semantic.

From his first presentation of it in 1967 up to his later writings about truth in the 90s, Davidson had relied on the Slingshot as his only weapon against the correspondence theory. But, perhaps as a consequence of the publication of Neale's book, Davidson has realized that the Goliath of the correspondence theory needs more than a single hit from a Slingshot — it needs the blast of a double-barreled shotgun. So in his final work in 2005, Davidson presented a set of considerations that amount to the Bradley Regress to show that no theory that accepts universals can account for the unity of the proposition or the fact. (On the other hand, that it *can* account for the unity of the proposition will be a key advantage of Davidson's own theory of truth.)

Here is the problem. The mark of the complete (assertive) sentence is that it's susceptible to truth and falsity in a way that a mere list is not. So the words in the sentence have to contribute something other than some set of particulars and universals. The contributions of the words have to be melded into something unified. Likewise the sentence itself must be a unity. If facts are left as mere sets, then the fact loses its unity; if the sentence is just a list, then the sentence loses its unity. Under these circumstances, facts are no longer able to serve as truth-makers and sentences no longer able to serve as truth-bearers. But if facts are integrated by some relation, such as the instantiation of the universal by the particular, or the co-instantiation of the two universals, then this relation must be another constituent of the fact, requiring integration with the other constituents just as they required integration with one another. Even the simplest facts, then, would have infinitely many constituents; an implausible view. Davidson explains:

> The problem is easier to state in semantic terms, and Plato gave us what we need to recognize it as a problem when he said that a sentence could not consist of a string of names or a string of verbs. The sentence 'Theaetetus sits' has a word that refers to, or names, Theaetetus, and a word whose function is somehow explained by mentioning the property (or form or universal) of Sitting. But the sentence says that Theaetetus *has* this property. If the semantics of the sentence were exhausted by referring to the two entities Theaetetus and the property of Sitting, it would be just a string of names; we would ask where the verb was. The verb, we understand, expresses the relation of instantiation. Our policy, however, is to explain verbs by relating them to properties and relations. But this cannot be the

> end of the matter, since we now have three entities, a person, a property, and a relation, but no verb. When we supply the appropriate verb, we will be forced to the next step, and so on. (Davidson 2005b, pp. 85-6.)

The problem is the double role of the predicate. On the one hand, the predicate is said to *contribute* a general entity, a universal. But so far, all we have is an inert list. The predicate must also somehow *attribute* the universal to the particular. These semantic tasks can be distinguished. We can reduce the predicate to one of its roles, the referential or contributive role, and allow some new word to attribute the universal to the particular. But that is now the predicate, and the original subject and predicate just a pair of subject terms. So the problem recurs.

At the outset, we may say that it's possible to contribute a universal to a sentence without attributing it to any particular. We may take 'Fa' and reduce the predicate to a mere contributor: 'a instantiates F.' Now 'F' brings F-ness to the sentence, but doesn't attribute F-ness to a. The new predicate 'instantiates' does that. Unfortunately, the new predicate also contributes a new universal, instantiation, and attributes it to the pair of a and F-ness. So, while we can have words that contribute universals without attributing them, it doesn't look as though we can have words that attribute universals without contributing them. For a word to do the job of relating the universal some other word contributes, it must contribute the relation between the universal and that to which it is attributed.

Nothing that we say about universals is really going to help with this problem. Frege may have come up with the least bad solution. For Frege, in 'Fa,' 'a' contributes the object a, and 'F,' we can misleadingly say at the outset, contributes the functional concept Fx. Concepts, for Frege, are functions that take objects as arguments and return The True or The False as values. If this can be made intelligible, we immediately account for the truth-aptitude of the proposition: a proposition contains an object and a function that takes that object and returns truth or falsity, so of course the proposition will be truth-apt. How do the functional concept and the object unify? Frege says that a concept "is

unsaturated — it contains an empty place; only when this place is filled up with a proper name, or with an expression that replaces a proper name, does a complete sense [thought, proposition] appear." (Frege 1891, p. 139) But *how* does such a thought appear? Functions don't take arguments and return values all by themselves: there must be some difference between just listing Fx and a, and giving a as an argument to Fx as a function.

For Frege, the difference is that listing contributions of different words is impossible, but using predicates attributively is possible. Frege tries to solve the problem by saying that 'F' *does not* contribute Fx (despite Fx's being its referent), it *only* attributes it: "A concept — as I understand the word — is predicative." (Frege 1892b, p. 182) and, "...the three words 'the concept "horse"' do designate an object, but on that very account they do not designate a concept, as I am using the word." (*ibid*, p. 184) So for Frege, a list of particular and universal (object and concept) isn't possible, since concepts can't be listed. Hence whenever a predicate appears, it appears in the attributive, never the contributive, mode.

Unfortunately, Frege finds himself in a conflicted position. On the one hand, he wants to tell us about concepts: they're unsaturated functions, and that's why they take objects as arguments and return truth-values. On the other hand, he wants predicates to appear only attributively. If we can't somehow use predicates to contribute concepts to the discussion, then we can hardly learn that they are unsaturated. But if we can somehow contribute them, then we can't use predicates attributively. Either we can't solve the problem because we don't know what we're talking about, or we can't solve the problem for the standard reason of the double role of the predicate.

Unfortunately, Davidson fails to understand just why Frege's approach won't work. He says, "...if, as Frege maintained, predicates refer to entities, and this fact exhausts their semantic role, it does not matter how odd or permeable some of the entities are, for we can still raise the question of how those entities are related to those other entities, objects." (Davidson 2005b, p. 145) The problem is quite the reverse: predicates

don't really refer to entities. Predicates' semantic role is exhausted not by *referring* to concepts but by attributing them to objects, but that leaves us in the dark about just what concepts are.

1.2.3 A Sophisticated Correspondence Theory

What we need is to divorce the contributive and attributive functions of the predicate. Herbert Hochberg, in proposing a more sophisticated correspondence theory, agrees:

> Thinking of truth values as values of a function [as in Frege] obscures a basic point. If one thinks in terms of facts, one easily sees that the function cannot be a constituent of the value, the fact, and also be what maps an argument onto such a value. The relevant constituent of a fact is a property or relation, while what maps a term onto a fact is a function. (Hochberg 2003, p. 79)

For Frege, the truth-maker for a sentence was The True, and the function that took the sentence and returned this truth-maker was the predicate. But that won't work. Hochberg suggests that the truth-maker is a fact. The function that takes the sentence and returns its truth-making fact is not the predicate; the predicate's role is to contribute a universal. The universal is a constituent of the truth-making fact, and is not a function that returns it. There is some other function that takes sentences and returns facts.

The predicate, then, has lost its attributive function. What constitutes the fact as a unity, then? Hochberg's Russellian answer is *logical form*:

> Forms like Φx are common forms of monadic atomic facts, but they are neither monadic properties of such facts nor dyadic relations obtaining between the "constituents" of such facts.... Facts can be said to be complex in that they have terms and attributes and are of a form. Thus the terms and attributes can be thought of as "components." But in that they are neither reducible to nor analyzable into such components, they are not complexes of them — and the form is not a component "in" the fact but a logical form of the fact — one can say "logical character of" but must not be misled into thinking that it is then a universal property exemplified by such facts. (*ibid*, p. 164-5)

The logical form of a fact is what unifies the referents of the subject and predicate into a unified entity capable of serving as truth-maker. But if we list out the contributions of

subject and predicate, we will still have exhausted the constituents of the fact. This is supposed to help solve the Bradley Regress by allowing facts to consist in finitely many elements while still being unified by an element that, though 'internal' to the fact, is also not a constituent in the fact.

Logical form, then, serves part of the attributive function, in holding universals together with the particulars that instantiate them. But attribution also has a linguistic aspect. Universals are attributes of particulars, independent of language, but we don't attribute universals to particulars, independent of language. We need some function from language to the world:

> The distinction between propositional functions, on the one hand, and properties and relations, on the other, allows us to arrive at one simple resolution of the Bradley-Frege problem about predication.... Let Φx be the logical form of a monadic atomic fact.... If we distinguish the form of the fact Fa, the form Φx, from the function (λΦ,x) Φx, where the latter, for the arguments F and a, has, as value, the fact Fa, we can recognize a further function that yields the same fact as value, for three arguments — the function (λΦ,x) Φx, the property F and the object a. But there is no further fact, or more complex form of a fact, involved.... (*ibid*, p. 81)

The λ-function is the language-world function we require. It takes sentences and returns truth-making facts.

Note that Hochberg repeatedly uses the phrase 'the fact Fa.' At a first pass we might guess that this is intended to name the truth-making fact for the sentence 'Fa.' But that won't work. Recall the Russellian argument against propositions from 1.1.1.3. If we can't identify an entity otherwise than by reference to its linguistic relata, we don't have an extralinguistic entity. But facts are intended to be extralinguistic. If 'the fact Fa' is intended to be shorthand for 'the fact which would make true the sentence "Fa" if it obtained,' then we have identified the fact with reference to its linguistic relata and don't, apparently, have an extralinguistic entity.

Logical form is not the only Russellian innovation to which Hochberg appeals. We may identify facts only by the use of definite descriptions, but the descriptions need

contain no linguistic relata of the facts. Allowing T(x,y) to mean that x is a term in y, A(x,y) to mean that x is attributed in y, and IN(x,y) to mean that x informs y, we may produce a definite description of the truth-maker for 'Fa':

ιp(T(a,p) & A(F,p) & IN(Φx,p))[19]

Having given the elements of the theory, let me now try to sum them up.

The Bradley Regress is the problem that predicates seem to have both attributive and contributive functions, but their contributive function "gets in the way" of their performing their attributive function. If the predicate contributes a universal, then it's not obvious how it also attributes the universal to the particular. The sophisticated theory divides what we took to be the function of the predicate into three. The contributive role is played by the predicate, which contributes a universal. The attributive role is divided into two, the extralinguistic role of uniting universal with particular in a fact, and the role of actively *using* the predicate to attribute the universal. The former role is played by logical forms of facts. The latter role is played by λ-functions that take sentences and return the facts in which the appropriate universals appear as attributes. No apparent regress threatens.

Nevertheless, the theory fails. First, we need to look for the λ-functions that relate language to fact. Consider the sentence 'Fa.' How does the λ-function "know" to return a fact that contains a and F and is informed by Φx? The latter doesn't seem to appear in the sentence, and the λ-function specifically takes *the sentence* 'Fa' and returns the appropriate fact. What in the sentence connects with the logical form? Hochberg explains, "Exemplification, as a logical form, is represented by sentential structure and the formation rules of the schema, and thus derivatively by 'Φx.'" (*ibid*, p. 165) This won't be sufficient.

Hochberg discusses a proposal by Sellars to eliminate the appearance of reference to relations by eliminating relational predicates, replacing them with spatial relations

[19] Hochberg 2003, p. 168

between subject *terms*; for instance, we might use 'a's being above 'b' to say that a loves b. Hochberg argues that this trick does not abolish universals:

> Sellars has created a notation where the predicates are not merely incomplete or unsaturated in Frege's sense [i.e., predicates like 'Fx' include space(s), in this case the 'x,' for the subject(s) and thereby display their incompleteness], but where they disappear when the individual variables or constants are "withdrawn." ...But this does not mean that we cannot isolate what plays the role of "R" in "aRb." It means that what plays the role of "R" cannot be set down, as I just set down a token of "R," without setting down tokens of subject terms. Sellars has failed to eliminate predicates. (*ibid*, p. 159)

Sellars's maneuver here seems to be exactly the same as Hochberg's. Sellars doesn't want for there to be a word to refer to a relation, so he employs structure to perform the task. As Hochberg points out, structure is just going proxy for the word; the word is still present but in a different, invisible, notation. Likewise: when Hochberg allows sentence structure to represent logical form, there is still a word for logical form present in the sentence, albeit in an invisible notation.

But assume that we reject this argument, and allow that sentence structure can represent logical form. We must refer to the fact independent of linguistic relata, and the structure of a sentence is linguistic relata. We must find, for theoretical purposes, some other means of referring to logical form. I digress and return to this issue below.

If logical form is contributed, then there seems to be a need to attribute it to the fact. Hochberg disagrees. He claims that form is internal to the fact, though not a constituent, and that the precise sense to be given to 'internal' will make clear how form can unify the fact without being a constituent. It would seem that the fact that the form informs the fact is a fact, which would require further analysis, thus reintroducing the Bradley Regress, but the sense in which the form's informing the fact is internal to the fact is intended to prevent this implication:

> ...that a is a term of such a fact is then expressed by the claim: a stands in T to [i.e., a is a term in] the fact having a as a term, F as attribute and Φx as form. And this is equivalent to saying that the fact exists, to "the fact having a as a term, F as attribute and Φx as form exists," on Russell's analysis of definite descriptions.

> Given that there is such a fact, it trivially follows that a is a term of it. For on Russell's theory we trivially have: ∃!(ιp)(T(a,p) & A(F,p) &IN(Φx,p)) ↔ T(a, ιp(T(a,p) & A(F,p) & IN(Φx,p)))).... The same is true of the logical relations A and IN.... In view of this feature of such relations, they can viably be taken to be... internal.... (*ibid*, p. 171)

It's true that the claim that the fact exists is materially equivalent to the claim that a is a term in it. So here are three biconditionals:

> ∃!(ιp)(T(a,p) & A(F,p) &IN(Φx,p)) ↔ T(a, ιp(T(a,p) & A(F,p) & IN(Φx,p)))
> ∃!(ιp)(T(a,p) & A(F,p) &IN(Φx,p)) ↔ A(F, ιp(T(a,p) & A(F,p) & IN(Φx,p)))
> ∃!(ιp)(T(a,p) & A(F,p) &IN(Φx,p)) ↔ IN(Φx, ιp(T(a,p) & A(F,p) & IN(Φx,p)))

But obviously these show a material equivalence between the three claims about the constituents and form of the fact, as well. But it doesn't seem plausible that a material equivalence between the claim that a is a term in a fact, and that F is attributed in it, suffices to show that they're true in virtue of the same fact. On the contrary, it seems perfectly plain that a's being a term in a fact and F's being an attribute in that fact are different facts. Likewise, Φx's being the form of the fact seems to be different from the fact itself. Φx's being the form of the fact was to be 'internal' to the fact in some special sense, but the only sense that I can see is material equivalence between the claim that the form informs the fact and the claim that the fact exists. Material equivalence is obviously insufficient for identity of truth-maker.

Hochberg points out[20] that the equivalence is an instance of a *Principia Mathematica* theorem, ∃!(ιx)(Φx) ↔ Φ(ιx)(Φx). Perhaps that is why the equivalence shows some sort of internality between the claims. The idea would be that each of the three claims that's materially equivalent with the existence claim for the fact will be internal to that claim, but not to each other. But this piece of *Principia Mathematica* is not so helpful. If we accept the theorem, we do so because we accept Russell's theory of descriptions. But on that theory, the materially equivalent claims are really the same claim in different notation; they're the same claim because they receive the same

[20] *ibid*, p. 171

canonical expansion into $\exists x[\Phi x \ \& \ \forall y(\Phi y \leftrightarrow x=y)]$. Thus, the material equivalence between the claim that Φx informs the fact, and the claim that the fact exists, is a matter of rephrasing. But this brings the informing of the fact and the fact into a closer relation than mere internality; it makes them identical. But if the fact that Φx informs the fact just *is* the fact, then it's not plain how Φx's informing the fact can account for the unity of the fact. There's no substantial relation between Φx's informing the fact and the fact any longer: we're appealing to the fact to account for its own unity. Only something external to the fact can substantially account for its unity; but only something internal to the fact can really unify it. To unify the fact, the form's being its form would have to be both external and internal.

Let me return, anti-climactically, to the matter of identifying logical form with reference to linguistic relata. We may refer to a and F-ness, presumably, with 'a' and 'F-ness;' with, in general, names and nominalized adjectives. But English, at least, seems to have no word for the logical form of a fact that includes a single term and a one-place relation. The previous sentence sought to refer to such forms with reference to the facts they unify. But our problem was to refer to facts, and we need to refer to forms to refer to facts. If we can refer to forms only with relation to facts, then our attempt to ground out reference is viciously circular.

We might try to refer to forms with reference to sentences that would correspond to facts with those forms, but this is patently to make forms linguistic, which they are not. Or perhaps we can point; Russell said:

> It is not at all clear what is the right logical account of "form," but whatever this account may be, it is clear that we have acquaintance (possibly in an extended sense of the word "acquaintance") with something as abstract as the pure form, since otherwise we could not use intelligently such a word as "relation."
>
> ...As a matter of introspection, it may often be hard to detect such acquaintance; but there is no doubt that, especially where very abstract matters are concerned, we really do have an acquaintance which we find it difficult to isolate or to become acquainted with. The introspective difficulty, therefore, cannot be

> regarded as fatal, or as outweighing a logical argument of which the data and the inference seem to allow little risk of error. (Russell 1913/1984, pp. 98-9)

Russell's argument is nicely put. He does not claim that we must be acquainted with *logical form* in order to grasp relations; he claims that we must be acquainted with something *as abstract as* logical form in order to grasp relations. He seems to be assuming that our objection to logical form is that they are abstracta, and we object to all abstracta. But the objection here is that logical form is inadequate to help solve the problem it was introduced to solve. I've accepted other abstracta, universals, for sake of argument, and I would accept logical forms, too, if only they could be made to work.

Given the apparent uselessness of logical form, Russell's second point, that the difficulty of finding acquaintance with forms should not persuade us that we have no such acquaintance, loses force. It's true that, if logical forms are the best explanation for our ability to entertain propositions, then we have good reason to believe that we have acquaintance with such forms, however difficult it may be to isolate the acquaintance. But logical form is not the best explanation; it is part of an inadequate explanation.

However, recall that our goal was to secure reference to facts independently of linguistic relata. The immediate problem is securing reference to forms independently of linguistic relata. The obvious solution is to refer to forms by indexical or name; by somehow attaching a referring term to the form. But whether forms exist or not, it's very hard to see how we could attach referring terms to them when we can't isolate our acquaintance with them. So it's not at all apparent that we can secure reference to forms, and hence to facts, in such a way that they're extralinguistic entities.

I conclude that the correspondence theory is hopeless. To avoid the Slingshot, we must posit universals. But universals bring on the Bradley Regress. To avoid this, we must posit logical forms. But these don't help, they just push off the Regress another step; and we might not be able to posit extralinguistic forms anyway. We can't have facts that both pass the Slingshot and also account for the unity of the proposition.

1.3 A DAVIDSONIAN APPROACH

1.3.1 The Importance of Truth-Conditions

Davidson is of two minds about the existence of truth-conditions. His official position is that there are no such things. Nevertheless, he thinks that a Tarskian truth theory of a certain kind can function as a theory of meaning. A Tarskian truth theory, one might think, specifies truth-conditions. If meanings are truth-conditions, and truth-conditions don't exist, what has happened to meanings? What do we give when we interpret?

It's in dealing with interpretation and the acquisition of language that Davidson's other view appears. Meanings in the guise of the contents of assertions appear repeatedly as tangible things when Davidson discusses how we interpret speakers and learn a first language. In this subsection, I compare the two positions and argue that Davidson needs truth-conditions. In the next two subsections, I argue that he can have them, too.

Davidson denies the existence of meanings rather brazenly at several places. In what follows, I will sometimes speak of how things would be if meanings were theoretical constructs (which, I contend, they are not). The phrasing is suggested by this passage: "...we must view meaning... as a theoretical construct. Like any construct, it is arbitrary except for the formal and empirical constraints we impose on it." (Davidson 1973, pp. 256-7) This is perhaps the clearest statement of the negation of meanings in Davidson's works, but it's a neat fit with only some of Davidson's other ideas. First I document the neat fits, then the poor ones.

Even before announcing the truth-conditional theory of meaning, Davidson had begun to attack meanings:

> Paradoxically, the one thing meanings do not seem to do is oil the wheels of a theory of meaning — at least as long as we require of such a theory that it non-trivially give the meaning of every sentence in the language. My objection to meanings in the theory of meaning is not that they are abstract or that their identity conditions are obscure, but that they have no demonstrated use. (*ibid.*, pp. 21-2)

The point is sharply put. Davidson does not argue that meanings do not exist: only that they are of no use for the theory of meaning. Here's what he means. If our theory of meaning is to be non-trivial, it can't just say that the meaning of "Φα" is the meaning of 'Φ,' somehow concatenated with the meaning of 'α.'

But Davidson goes further than to announce the uselessness of meanings for the theory of meaning. He denies the existence of meanings: for by denying the existence of truth-conditions, while maintaining that truth-conditions are meanings, one denies meanings. And he does deny the existence of truth-conditions:

> Nothing…, no *thing*, makes sentences and theories true: not experience, not surface irritations, not the world, can make a sentence true. *That* experience takes a certain course, that our skin is warmed or punctured, that the universe is finite, these facts, if we like to talk that way, make sentences and theories true. But this point is put better without mention of facts. The sentence 'My skin is warm' is true if and only if my skin is warm. Here there is no reference to a fact, a world, an experience, or a piece of evidence. (Davidson 1974, p. 194)

Here, Davidson claims that "that experience takes a certain course," "that our skin is warmed or punctured," "that the universe is finite," make no reference to facts, the world, experiences, or pieces of evidence. To what *do* "that" clauses make reference? According to Davidson[21], the answer is *sentences*, but that won't do. *The sentence* "Our skin is warmed" does not make it true that our skin is warmed. Rather, *our skin's being warmed* makes it true that our skin is warmed, but our skin's being warmed is not a sentence.

Elsewhere, in the paper "On Saying That," Davidson makes another move that requires truth-conditions. Davidson considers several theories of belief attribution, finally settling on one that makes use of his paratactic analysis of "that" clauses. For Davidson, a sentence like, "Galileo said that the earth moves" is to be analyzed like this: "The earth moves. Galileo said that." 'That' in the second sentence refers to the first sentence itself. But of course, Galileo didn't say "The earth moves." He didn't speak English. So how can 'that' refer to what Galileo said, when it doesn't refer to any utterance of Galileo's? It

[21] See Davidson 1968.

seems to me that 'that,' in "Galileo said that," is ambiguous between two readings. On one reading, it refers to an actual utterance by Galileo. On another, it refers to the meaning of that utterance. Davidson would disagree. He remarks:

> We are indeed asked to make sense of a judgement of synonymy between utterances, but not as the foundation of a theory of language, merely as an unanalyzed part of the content of the familiar idiom of indirect discourse….
>
> The fact that an informal paraphrase of the predicate ['samesays;' x and y are samesayers just in case x says something and y says something with the same meaning] appeals to a relation of sameness of content between utterances introduces no intensional entities or semantics. Some have regarded this as a form of cheating, but the policy is deliberate and principled. [There is a] distinction between questions of logical form (which is the present concern) and the analysis of individual predicates…. It is also worth observing that radical interpretation, if it is succeeds, yields an adequate concept of synonymy as between utterances. (Davidson 1968, p. 104; content after first ellipsis is from footnote 14, same page)

The passage is complicated. First, Davidson (twice, in different words) distinguishes between introducing semantic entities, in this case meanings, when doing semantics, and introducing them when analyzing some particular predicate. Second, he asserts that radical interpretation will ground a concept of synonymy; presumably one that does not require the existence of meanings.

The first point we can accept with equanimity. In "On Saying That," Davidson can narrow his concern to the logical form of "that" clauses, including ones in belief attributions. That task doesn't require the introduction of meanings: it is a semantic task, and, ironically, meanings don't help with theory of meaning. The second point is more doubtful. The only relevance of the claim that radical interpretation will give us all the synonymy we need is to deny that we will need meanings even when we do come to analyzing the 'samesays' predicate. But it is false to say that radical interpretation does not appeal to meanings. On the contrary: at every step in his discussions of radical interpretation, Davidson appeals to meanings. The pattern is massive and overwhelming, if also, at times, obscure.

Before I turn to that pattern, I need to deal with two problems. First, in section 1.1.3, I used a Russellian argument against propositions. The argument ran as follows. Meanings, Fregean senses, are not to be thought of as linguistic entities. But, if an entity can only be identified with reference to its linguistic relata, then it seems to be a linguistic entity; indeed, being identifiable only with reference to linguistic relata might even be definitional of the linguistic nature. But Fregean senses are only identifiable with "that" clauses in which 'that' takes its reference from the sentence, a linguistic entity, that follows it. Hence Fregean senses are linguistic entities, though they're not supposed to be.

The problem is that I seem to have replicated the analysis, except that my meanings are truth-conditions, not senses. Thus the same argument should apply to my view. Nevertheless, it doesn't.

Fregean meanings and truth-conditions are very different sorts of things. Fregean senses play no causal or explanatory role in the world; their role is exhausted by their being invoked in semantics and interpretation. But at least some truth-conditions are tangible things: events with causal relations. They can thus be identified otherwise than with reference to linguistic relata: they can be identified with reference to their causes or effects, and in many cases by pointing. For instance, one might say that "*That*," uttered while pointing, "makes it true that there is an explosion (at that place at that time)." The first instance of 'that' refers to the indexed object, the explosion. The second instance of 'that' refers to the sentence, "There is an explosion." The first instance of 'that' thus refers to the truth-condition of the referent of the second instance of 'that.' The truth-condition has thus been identified otherwise than with reference to its linguistic relata. Of course, it would be easy to refer to it with reference to linguistic relata: in "The truth-condition of 'There is an explosion' obtains," the truth-condition is referred to solely with reference to the sentence whose truth-condition it is. But that we can identify a truth-condition with reference to language doesn't imply that we can't identity it otherwise, and the objection

to Fregean senses was that they could not be identified otherwise than by reference to language, not that they could be identified with reference to language.

The second problem I should deal with has to do with the paratactic analysis. In section 1.1.4, I appealed to Davidson's paratactic analysis when attacking minimalism. But here I have claimed that the paratactic analysis is not right. The paratactic analysis is of limited applicability, or is incomplete.

Consider this sentence: "When I say that the cat is ill, I say that because it matters that the cat is ill." Here, the paratactic analysis works for two instances of 'that;' it requires modification for the other. Consider how things would be if the paratactic analysis were correct. Since the referent of all three 'that's seems to be "the cat is ill," we could replace the sentence with this:

The cat is ill.

When I say that, I say that because that matters.

But this pair plainly has a different meaning from the original sentence. If all three 'that's refer to the preceding sentence, the sentence is false. The problem is that *the sentence* "The cat is ill" doesn't matter (according to me), so I don't say anything because of its' mattering. I say that the cat is ill because of something else, something that does matter: the cat's being ill. The cat's being ill is not a sentence, but some other kind of thing. It's to that something else that I refer when I say that it matters that the cat is ill.

The paratactic analysis, then, is incomplete. Some instances of 'that' that appear in "that" clauses do refer, as Davidson says, to utterances or sentences. But others refer to things independent of language: things that matter, have causal power, and so forth.

This revision does not fall afoul of the problem of compositionality. On an analysis of "that" clauses according to which "that the cat is ill" is to be treated as a unified singular referring phrase, we must learn a new linguistic primitive every time we hear a new "that" clause. Here, though, the word 'that' is the linguistic primitive, and it refers to some entity connected with the sentence that follows the word 'that.' The

sentence, of course, does not refer to that entity or any other; sentences are not referring terms. Since the problem of compositionality would emerge only if the "that" clause were itself a referring device, and it is not, compositionality requirements do not block my approach.[22]

Why does Davidson deny that there are truth-conditions? The Slingshot provides the narrowest reason. In 1.3.3, I'll deal with the Slingshot. But there is a broader reason, having to do with an imagined connection between traditional correspondence theory, representationalism, and realism:

> …the real objection [to correspondence theory] is… that such theories fail to provide entities to which truth vehicles can be said to correspond. If this is right, and I am convinced it is, we ought also to question the popular assumption that sentences, or their spoken tokens, or sentence-like entities or configurations in our brains, can properly be called "representations," since there is nothing for them to represent. If we give up facts as entities that make sentences true, we ought to give up representations at the same time, for the legitimacy of each depends on the legitimacy of the other. (Davidson 2005b, p. 41)

He continues:

> The realist view of truth, if it has any content, must be based on the idea of correspondence, correspondence as applied to sentences or beliefs or utterances — entities that are propositional in character; and such correspondence cannot be made intelligible. …it is futile either to reject or to accept the slogan that the real and the true are "independent of our beliefs." The only evident positive sense we can make of this phrase, the only use that consorts with the intentions of those who prize it, derives from the idea of correspondence, and this is an idea without content. (*ibid*, pp. 41-2)

A sentence is said to correspond to whatever it represents; that is realism. But we cannot make sense of correspondence, so we cannot make sense of representation: so realism is an empty slogan.

We have a fairly good idea of the problem Davidson has with correspondence. Rejecting traditional correspondence theories that appeal to facts of Russellian vintage,

[22] To my mind, this counts as a revision, not a rejection, of Davidson's account. Davidson's insight was a trick allowing us to deny that "that" clauses were, as wholes, referring terms, and my revision accepts the trick. It rejects the claim that every referent of an instance of 'that' that begins a "that" clause refers to a sentence. But whether I've "revised" or "rejected" Davidson's account is a semantic issue of the sort that even philosophers should ignore.

he thinks that he has rejected any theory that makes language represent reality. But the narrowness of the traditional correspondence theories is out of tune with the breadth of the notion of linguistic representation. Rejecting the one does not require rejecting the other.

Davidson's conception of representation is never made clear. But we can tell that whatever it is for x to represent y, it's for realism to be true about x, and realism's being true about x is sufficient for skepticism about the truth of x:

> [Realism accepts] the intuitive idea that truth, aside from a few special cases, is entirely independent of our beliefs; as it is sometimes put, our beliefs might be just as they are and yet reality — and so the truth about reality — be very different. (*ibid*, p. 33)

Realism might sometimes be put this way, but probably not by realists. Nobody adopts realism to ground the case for scepticism. We adopt realism because we think that the mind is hooked up to a reality beyond itself. The fact that traditional realism of the Descartes-Russell variety never figured out how the hook-ups worked notwithstanding, the idea of realism was never that we were out of touch with external reality, it was that there is an external reality with which to be in touch.

We can have realism without scepticism. Realism is only the claim that our sentences, when true, are true because of their meanings and the way things stand with meant reality. We can have linguistic representation, too: a sentence represents as obtaining the condition under which it would be true. It doesn't represent by resembling, and there's no sense at all to be made of comparing the two, as sense could be made of comparing Cartesian representations to what they represent. And it doesn't represent by naming: a sentence is not a name for a fact.

Davidson's denial of truth-conditions is just the denial of facts coupled with the failure to grasp that there could be any form of realism other than Cartesian or Russellian. This failure is a colossal failure of Davidson's usually colorful philosophical imagination.

Because of his official position, we won't really catch Davidson claiming that truth-conditions exist. But we'll find him coming so close that it's hard to see what *else* he might have meant. The interpretations to follow rest on some equations: that meanings are truth-conditions, and that anything called the 'content' or 'subject matter' of a sentence is its meaning.

It might seem that with those equations in place, it would really be impossible for Davidson to deny the existence of truth-conditions. Of course sentences have content, so if I find him saying that sentences have content and decide that I've found him saying that truth-conditions exist, then it might seem that I've cheated. I need to be clearer about what I mean. *Of course* truth-conditions exist, in some sense. What I want to point out is that, for Davidson, truth-conditions don't exist just as theoretical constructs: they're concrete entities that exist independently of language and independently of semantic theories of truth for languages. So what I'm looking for is not passages that just have it that truth-conditions, meanings, contents, or subject matters exist, but that identify any of those things with external situations or events that don't seem like theoretical or linguistic entities.

Here is a fairly clear statement that makes subject matter exist:

> My approach [to interpretation] is externalist: I suggest that interpretation depends (in the simplest and most basic situations) on the external objects and events salient to both speaker and interpreter, the very objects and events which the speaker's words are then taken by the interpreter to have as subject matter. (*ibid*, p. 64)

Here, external objects and events are taken to be subject matter. Interpretation, that is, the giving of meanings, depends on subject matter. That's obvious: subject matter is meaning, and of course giving meanings depends on meanings. What's relevant here is that meanings are held to involve concrete objects and events. These are not influences among theoretical entities (as meanings might be if they were non-existent truth-conditions that "existed" only in the structure of a Tarskian theory), but real, tangible things that exist. Perhaps clearer:

> One place to begin is by asking how the sentences directly tied to perception get their content. ...Perceptual sentences have an empirical content given by the situations which stir us to accept or reject them, and the same goes for the beliefs expressed by those sentences. (*ibid*, 1997a. p. 137)

It's hard to see what it would be for a situation to "give" a sentence empirical content, unless it amounts to *being* that empirical content. To be sure, Davidson continues to point out that more is required for a sentence to have a given empirical content than that it is held true under the right circumstances: it must be embedded in the right context of other beliefs. For instance, I can be trained to say "There's an electron" whenever the cloud chamber shows one, without having any idea what an electron is. In that case, it's plausible to say that the empirical content of the sentence is really "There's a little line-shaped cloud." But that's not to say that there's any empirical content to "There's an electron," spoken by the competent physicist, other than the situation that prompts the utterance.

It's important that perceptual beliefs are fairly basic. Whatever it is that prompts my assent to "$e=mc^2$," it's surely not some specific situation or event — at least not if that utterance on my part is to have its customary content. The truth-condition of that utterance is no particular happening in the world, but something of an entirely different order. Perceptual beliefs have easily recognizable truth-conditions, but many other beliefs have much more exotic truth-conditions.

In arguing for the claim that language is public (without resting on the false premise that language is necessarily conventional), Davidson says:

> Without... sharing of reactions to common stimuli, thought and speech would have no particular content — that is, no content at all. It takes two points of view to give a location to the cause of a thought, and thus to define its content.... [When a] common cause has been determined... the triangle which gives content to thought and speech is complete. (*ibid*, pp. 212-3)

Here, giving location to the cause of a thought is defining its content. But giving location seems very much like defining, so there's excellent reason to believe that cause is content. Relatedly and with perfect clarity, Davidson says that:

> ...we must, in the plainest and methodologically most basic cases, take the objects of a belief to be the causes of that belief.... Communication begins where causes converge: your utterance means what mine does if belief in its truth is systematically caused by the same events and objects. (*ibid*, 1983. p. 151)[23]

Can the object of a belief just be the referent of the subject of the sentence that expresses the belief? It cannot. If the belief that Fa had its content just because a caused it, then there would be no difference between the content of Fa and the content of Ga. It's a's *being F* that is the content, hence object, of the belief, so it's a's being F that is the cause of the belief (in the plainest and most methodologically basic cases). But if the truth-condition is the cause, then, since causes are concrete things that exist in a pretty ordinary sense, then at least some truth-conditions are ordinary entities, not strange theoretical abstracta.

Finally, Davidson remarks: "If anything is systematically causing certain experiences (or verbal responses), that is what the thoughts and utterances are about.... If nothing is systematically causing the experiences, there is no content..." (Davidson 1990, p. 201). Here, the (typical) cause is the content. But content is surely the same as meaning, and meaning is truth-condition. Hence, for at least some assertive utterances (perceptual sentences), the truth-condition is the event that causes (prompts) the utterance. These truth-conditions are tangible, fully really, concrete things. (Of course, this account needs to be enriched for the many cases of non-perceptual sentences.)

In this subsection, I've reviewed Davidson's rhetoric against truth-conditions and found that his view was deeply confused. I've shown that he needs truth-conditions to be the referents of some instances of 'that,' and that he seems to appeal to concretely existing truth-conditions in discussions of interpretation. There are *many more* passages that I could cite with similar import to the last few. It remains to be seen that we can have the concept of a truth-condition, but Davidson himself has given us no reason more general

[23] Davidson is so pleased with this formulation that he quoted himself in the 1990 paper "Epistemology Externalized," a paper that postdates his "Afterthoughts" on "A Coherence Theory of Truth and Knowledge."

than the Slingshot, and has given a view that seems to require there to be truth-conditions.

1.3.2 The Transcendence of Truth and the Necessity for Truth-Conditions

It is impossible to give a theory of truth; it is impossible to give a theory of truth-conditions. It is possible to say a few informative things about both, but the concept of truth cannot be articulated.

If we can't have a theory of truth, then what are Tarski's semantic theories? Semantic theories are not theories of truth, but theories of truth in languages. Semantic theories don't even *attempt* to articulate the concept of truth: on the contrary, one must already have the concept of truth before one can even figure out what's going on with semantic theories:

> The central difficulty [with the thought that a semantic theory is a theory of truth] is due simply to the fact that Tarski does not tell us how to apply the concept to a new case, whether the new case is a new language or a word newly added to a language.... This feature of Tarski's definitions can in turn easily be traced to the fact that they depend on giving the extension or reference of the basic predicates or names by enumerating cases: a definition given in this way can provide no clue for the next or general case. (Davidson 2005b, p. 17)

Tarski's theories just list the truth-conditions of all the sentences of a language; or, more precisely, derive that infinite list from a set of axioms of reference and satisfaction. There is no metatheory from which the *axioms* are derived; nor does a set of axioms work for more than one language.

To be recognizable as theories of truth for languages, Tarskian theories must be recognized by someone who grasps what truth in a language is, and that can only be grasped by someone with the concept of truth. Without an extra- or at least inter-linguistic concept of truth, one is ill-placed to grasp the *point* of Tarskian theories: what it is that they're all theories of, what holds them together. That is why minimalism in which the truth-bearers are sentences is a hopeless theory. The minimalist claims that the list of T-sentences exhausts the content of the concept of truth, but one has to know more than

the list to grasp just what the list is. The concept of truth transcends the articulate theories offered by Tarski and minimalists.

The concept of truth has empirical significance and ties to the concepts of belief and desire. None of that is captured by Tarskian theories. Consider the empirical significance. Davidson explains:

> ...given a language we understand, an interpreted language such as English, we recognize as true all sentences of the form "'Snow is white' is true if and only if snow is white." Tarski calls such sentences "partial definitions" of truth. Obviously, a definition that entails all such sentences will have the same extension (for the specified language) as the intuitive concept of truth with which we started. To admit this is to count T-sentences as having empirical content; otherwise Convention-T would have no point, nor would Tarski's insistence that he is interested in defining truth only for interpreted languages. (*ibid*, p. 23)

Tarskian truth theories, on their own, appear to be mere stipulations. But they are intended to be responsive to empirical facts about what speakers mean. Davidson makes this point in two ways. First, if Tarski's definitions were mere stipulations, then why would Convention-T be a requirement on such definitions? Convention-T tells us that a theory of truth for a language is to count a sentence as true for that language only if that sentence is true. But whether the sentence is true is an empirical fact about that sentence's meaning and how things stand with that meant reality.

Second, if Tarski's definitions were mere stipulations, then why would Tarski intend for them to apply to interpreted languages? Any theory that tries to state the truth-conditions for sentences in an interpreted language is obviously responsible to the truth-conditions that those sentences actually have. This empirical basis for Tarski's theories is obviously not captured by those theories; this is one way in which the concept of truth outstrips what we can say about it.

The obvious reply is that the concept of truth outstrips what *Tarski* has said about it, but that we need to say more. One might entertain the notion of introducing a theory of truth-conditions — for instance, a theory of facts — and articulating the concept of truth with reference to those concepts. Davidson remarks: "It is a mistake to look for an

explicit definition or outright reduction of the concept of truth. Truth is one of the clearest and most basic concepts we have, so it is fruitless to dream of eliminating it in favor of something simpler or more fundamental." (*ibid*, p. 55) There is something on face absurd about trying to define truth. To give a definition is to clarify a concept, by giving its — as yet ungrasped — meaning in terms that are already fully grasped. But what concept do we grasp before truth? And what philosophical construct will be more clear?

To grasp any sentence at all, it's necessary to grasp its truth-conditions; that is, the circumstances under which it is true. That claim will be true on any reasonable theory of meaning. The truth-conditional theory of meaning goes further and says that grasping truth-conditions is all that there is to grasping meanings, but every theory will agree that if you don't know the circumstances under which a sentence is true, you don't understand the sentence. But if it's impossible to understand any sentence without knowing the conditions under which it is *true*, an implicit grasp of the concept of truth is a prerequisite for understanding *anything* that is understood in the form of a sentence. Thus literally *nothing* can be grasped prior to a grasp of the concept of truth.

Likewise, surely no philosophical notion introduced to help articulate truth could possibly be clearer than the concept of truth. As the discussion of sophisticated correspondence theory should have shown, the notion of logical form, or something of the sort, is a prerequisite on any theory of facts that is to survive the slingshot. But surely the notion of logical form is no clearer, no more obvious, than the concept of truth? No piece of philosophical artifice can help us articulate the concept of truth: the philosophical invention will of necessity be less familiar to us than the concept of truth. That's not to say that we can't use philosophical notions to say interesting things about truth, only that they can't make the concept of truth any clearer or more understood than it already is.

If truth can't be articulated, neither can the notion of a truth-condition. Davidson's complaint that correspondence theories "fail to provide entities to which truth vehicles

can be said to correspond" is odd. To articulate the concept of truth-conditions, one would have to articulate the concept of something that determines whether a sentence is true or false. Surely we have a handle on determination. So to articulate the concept of truth-conditions, we would have to articulate the concept of truth, which we cannot do. So the problem with correspondence theories is not that they *fail* to provide truth-conditions, it's that they *try* to provide truth-conditions.

The concept of truth, and with it the concept of a truth-condition, are unanalysably fundamental to the operation of thought. That permits us a certain looseness in talking about truth-conditions. No theorist can be held responsible for producing an exhaustive theory of truth-conditions, since no such theory can be produced. We can say many interesting things about truth-conditions; for instance, that some of them are events, that some of them probably are not events, and so forth. But we can't say much that will be true of all truth-conditions.

Given the centrality and ubiquity of the concept of truth, a certain argument of Friedrich Hayek's would make it seem obvious that the concept can't be articulated.[24] Hayek argues that it's inherent in the phenomenon of representation that no system can represent itself. For x to represent y, there must be an isomorphism of some kind between the elements of x and the elements of y. But for x to represent at all, it must be embedded within a larger system, within which it takes on representational power. With no such larger system, a so-called representation would just be a physical object with no intentionality. Hence, for x to represent y, x must be embedded in a system more complicated then y. So for a system to represent itself, it must be more complicated than it is: it must be as complicated as it is to have the relevant isomorphism, but more complicated than that, in order to have intentionality at all.

[24] See Hayek 1952, pp. 116, 179-86, Hayek, 1955, pp. 8-11, Hayek, 1963, p. 60, Hayek, 1964, p. 25. see also Register, 2003, pp. 9-35, for a critical synthesis of these texts.

Since the concept of truth is so central to our mental lives — we can't understand language, and hence thought, without it — and is so ubiquitous — each and every understood sentence or thought displays a unique application of the concept of truth — to articulate the concept of truth seems tantamount to articulating one's own mind, and this one cannot do.

Of course, many systems can represent themselves with some degree of coarse-grainedness. That's why we can say anything informative about ourselves, and about the concept of truth. But a general theory of truth will run into one or another limit. It either fails of real generality, and is really just a set of true claims about truth that apply in specific contexts, or else attains generality at the cost of illumination and applicability. My present approach is not to try for a general theory of truth, but to limit myself to some remarks about truth (truth-conditions) in specific contexts.

1.3.3 The Unity of the Proposition

If we accept the theory of descriptions, Neale says, we can have a correspondence theory that will survive the slingshot. The problem is this. The theory of descriptions is a substantive piece of semantics. If we accept Tarski's conception of semantics — that it characterizes the connections between language and world — then there should be some sort of metaphysical analogue to the theory of descriptions, something that accounts for the difference of meaning between co-referential definite descriptions. This metaphysical analogue will not, of course, be an entity of any kind. It will be a match in structure between our metaphysics (of truth, say) and our semantics (the theory of descriptions). The theory of universals is one such account. By providing universals to be contributed by the predicates that appear within the description, it explains why the description contributes, not its referent, but one or more universals and some quantificational structure to the sentence in which it appears. Universals, sadly, led to the decomposition of the proposition. If we want to survive the slingshot, then, we need some different

metaphysical analogue for the theory of descriptions: some different account of the work done by predicates. As Davidson says:

> [Neale] shows in convincing detail how awkward it is to evade the argument. It can be done, as Russell's semantics did it, by making properties parts of facts and so the entities that correspond to predicates. This is a course I have argued against on the grounds that it cannot be incorporated into a satisfactory theory or definition of truth, and entities that are made up in part of abstract entities can hardly be thought of as empirical truth-makers. (Davidson 1999, p. 667)

But it can be done otherwise than by introducing universals. Davidson would go on to provide just the trick necessary — or rather, Davidson would go on to point out that Tarski had already provided just the trick necessary, without anyone noticing.

The problem of the unity of the proposition is the problem of the double function of the predicate. The predicate must attribute something to the subject. But, it has been thought, it must then contribute whatever it attributes to the subject. But it's not clear how a predicate can both refer and predicate. A naïve view might ignore the problem. A sophisticated view might introduce some further element, like logical form, to do the attributive work, while the predicate does the contributive work.

The best approach, Frege's, had the contributive function lying around more or less as an afterthought. The predicate attributes a concept to the subject, but never contributes the concept. Unfortunately, the account of the truth of the proposition relied on the concept, which was a function taking referents of subjects as arguments and returning True or False as values. To solve the problem, we need to go beyond Frege. He was right to say that predicates refer to nothing, but wrong to introduce concepts.

Frege's near approach to the best view is consistent with Frege's following his dictum that, "The meaning of a word must be asked for in the context of a proposition, not in isolation." (Frege 1884, p. 90) Frege begins with the proposition, the truth-bearer. Outside of the context of a proposition, the predicate has no meaning at all; inside that context, its meaning is its functional role of returning True or False given a reference.

That is, the semantic role of the predicate is specifically what's necessary to account for the truth of the proposition.

Why should things come out so neatly? On Davidson's account, the sentence and its truth-value is the given, and words are theoretical entities that we posit to account for the truth or falsity of sentences:

> ...what is most open to observation is sentences and their uses, and truth is the semantic concept we understand best. Reference and related semantic notions like satisfaction are by comparison theoretical (as are the notions of word, predicate, sentential connective, and the rest); there is no question about their 'correctness' beyond the question whether they yield a satisfactory account of the truth-conditions of sentences and the relations among sentences. (Davidson 1988, p. 181)

The semantic concept with which we begin is the concept of truth, and to give a theory of meaning for a speaker is to give a theory of truth for his utterances. But why should such a theory have any substructure at all? Why should it do anything but give the infinite list of T-sentences? Since truth-conditions are meanings, we do not understand sentences unless we know their truth-conditions, which are given in T-sentences. But because our minds are finite, and sentences are infinite in number, we need some finite theory from which we can derive the infinite list of T-sentences. For that reason, we need the axioms of a Tarskian theory of truth for a language, which are finite in number but adequate for the derivation of the T-sentences. These axioms assert the existence of theoretical entities: words. They divide these words into kinds: quantifiers, predicates, and so forth. They then attribute properties to these theoretical entities: this entity has the property of referring to this object; this entity has the property of being satisfied by these objects, and so forth. The predicates have exactly the features — satisfaction, mainly — necessary to account for the truth-value of sentences because they are theoretical posits designed for the purpose of so accounting.

It might seem that an apparatus designed in this way would have to be trivial. The predicates can't explain what they're introduced to explain, because their contribution is

defined with reference to what it is to account for. This objection misses its mark. The overall concept of predicate satisfaction is devoid of explanatory value; the fact that it applies to a word is constitutive of that word's being a predicate. But that's a matter of course. Concepts don't explain things: sentences explain things. Particular claims about which objects satisfy which predicates do have explanatory value, and those are some of the claims made in a Tarskian theory. The correct solution is this:

> Tarski's essential innovation is to make ingenious use of the idea that predicates are *true of* the entities which are named by the constants that occupy their spaces or are quantified over by the variables which appear in the same spaces and are bound by quantifiers. (Davidson 2005b, p. 159)

A predicate need make no contribution whatsoever. The predicate itself is what's satisfied by the subject; the attribution function, disappearing into the function of being satisfied, is the sole function of the predicate. The predicate's role of being satisfied perfectly accounts for the unity of the proposition. It could do nothing else.

There is an important objection to this idea, captured clearly by Hochberg:

> To avoid taking "F" to represent an attribute, and the sentence to represent a fact, Davidson talks in terms of "F," or "Fx," being satisfied by, or, as Quine sometimes puts it, *true of* a. ...one can treat satisfaction as a relation between a thing or sequence and a sign (predicate, open sentence) and take the truth maker as a dyadic fact that has a linguistic item, a predicate or open sentence, as a constituent. Thus we arrive at a variant of linguistic idealism, in that a linguistic item is involved in the truth maker, rather than merely the truth bearer. (Hochberg 2003, p. 176)

Since the whole thrust of this dissertation is intended to be in the direction of realism, a conviction of idealism would call for capital punishment. Hochberg takes for granted that Davidson's view of truth is exactly like the classical correspondence view, except with predicates replacing universals. But the object's satisfying the predicate does not exhaust the truth-maker. Rather, truth-conditions are non-linguistic entities. If something — an event, say — that could have been a truth-condition turns out to actually be one — there is an utterance that is true just in case that truth-condition exists — then the event could

be described as an object's satisfying a predicate. But, even in the absence of the predicate, the same event would have existed and had most of its causal relations intact.

With that in mind, let me return to Hochberg's argument. The best way to characterize the difference between linguistic realism and idealism is counterfactually: as a matter of what would have been, had things with language been different. Assume that Fa. But then, even if 'F' had not been coined, it still would have been the case that Fa (except in some unusual cases).

The counterfactual needs some accounting for, because it has a "that" clause in it, and, as I've argued, some such clauses are systematically ambiguous. What have I said is the case? What is "that Fa"? Does 'that' in this "that" clause refer to the sentence that follows, or to its truth-condition? The following is false:

Had 'F' not been coined, 'Fa' would have been true.

That captures the linguistic reading by quoting the referred-to sentence. But this sentence is false because, had 'F' not been coined, 'Fa' would have been meaningless. But this is true:

Had 'F' not been coined, it would have been the case that Fa.

Here, it's plain that 'that' does not refer to 'Fa,' but to 'Fa's truth condition, which is (in most cases) extralinguistic.

In addition to unifying the proposition, this conception of predication also gives the metaphysical analogue of the theory of descriptions. Instead of contributing an object, a definite description is an "incomplete symbol" that, in context, asserts that some unique object satisfies one or more predicates. The account of predication accounts for the semantic contribution of definite descriptions.

The theory of descriptions is a semantic theory, which has a limited ontological analogue: the account of predication. The account of predication is not a general theory of truth conditions, so it is not a general theory that makes truth-conditions into linguistic or partly linguistic entities.

The present perspective allows us to see more deeply what's wrong with traditional correspondence theories. The traditional notion of a fact, composed of particular, universal, and unifying form, tried to analyze *away* the concept of truth, by replacing it with concepts like reference. But the concept with which we must begin is the concept of truth; reference can only be understood as a theoretical property of words, which themselves must be understood as theoretical entities introduced to help account for the truth of sentences. By insisting that the parts of a sentence all refer to something, the traditional fact-theorist wants to look for the meaning of a word outside of the context of any proposition; that cannot succeed.

The fact that Davidson's account of predication provides the ontological analogue of the theory of descriptions, and so, like that theory, permits a theory of truth that involves truth-conditions, earns for Davidson the right to have truth-conditions. His paratactic account and his accounts of radical interpretation required just those, so I think that Davidson's view is substantially cleaner with truth-conditions added explicitly.

1.3.4 How Events Can be Truth-Conditions

On the face of things, there's no problem with events being truth-conditions. Plenty of unusual, but still real, objects have been thought of as truth-conditions. Mulligan, Simons, and Smith, in a justly well-regarded paper,[25] have suggested that tropes ("moments") are truth-makers. If it's even remotely plausible that an instance of a color can make a sentence true, then there's no presumption against an event making a sentence true.

Nevertheless, some philosophers have argued that no event is a fact — where "fact" just means truth-maker or truth-condition. In this section I deal with the arguments of Jonathan Bennett. Bennett accepts that events exist, but claims that they supervene on facts, and that no event is a fact. He makes two arguments to this effect. The first is a

[25] Mulligan, Simons, and Smith, 1984.

linguistic argument, directed against Davidson's account of the semantics of event-talk. The second is an argument dealing with the nature of event-causation.

Davidson argues that many sentences quantify over events, and hence introduce events into the ontology of the speaker. For instance, "He gesticulated wildly," if it is to imply that he gesticulated (without the addition of the premise that everything that gesticulated wildly, gesticulated), must be read as something like, "There was an event such that he performed it, it was a gesticulation, and it was wild." Since the account strives for generality, "He gesticulated" will have the same reading, less the final clause that mentions the wildness of the gesticulation. Obviously the conjunction with which we parsed the first sentence implies the second sentence.

Bennett does not try to undercut Davidson's argument for the conclusion; rather, he confronts the conclusion head-on. Davidson's view gives *reference* to, and *quantification* over, events primacy over *attributing* events, or participation in them, to objects. This is grammatically problematic, according to Bennett:

> Do we understand "He gestured" *through* the thought that he made a gesture? If Davidson's theory is correct, we do, and then one of the following must be true: (1) We could not educate a child into knowing a big fragment of English from which perfect nominals [which phrases, Bennett argues, we use to refer to events] were absent.... (2) We could do it, but the child would perform clumsily with adverb-dropping inferences, not having our smooth rules for handling them.... (3) The child would think differently from us, or perform worse than us, because he would employ the perfect-nominal apparatus in his thoughts: he would *think* "They performed well" in the form "Their performance was good", even though he could not *say* the latter. (Bennett 1988, p. 18)

In a spirit of charity, we should overlook the obvious reading, which makes the argument blatantly question-begging. The obvious reading has it that, since Davidson is wrong about us, if his theory applied to someone, the someone would speak differently from the way we do, which we do not: so his theory does not apply to us. Hopefully, something deeper is going on. One could read the passage as a somewhat occulted challenge: Davidson, explain to us why, if your theory is true, we speak as we do. The challenge

then takes the form of a trilemma: choose one of these three disastrous consequences of your theory, Davidson, and make it seem less disastrous.

Even so, the second consequence is unsalvageable as a challenge to Davidson. There is no way to take it as anything other than a question-begging assertion. If Davidson's theory is true, then we have the smooth rules for adverb-dropping inferences that we think we have — indeed, the point of Davidson's theory is to account for just those smooth inferences, and Bennett gives no reason to think that Davidson's theory fails at this point. So really there are only two disasters to choose from.

Consider the third point: that if Davidson's theory were true, we would think "Their performance was good," rather than "They performed well." The difference between these two sentences, from my Davidsonian perspective, is in their degree of explicitness. The first one explicitly quantifies over an event: it says that there was an event that was a performance, it was by them, and it was good. But the second one implicitly says just the same. The second one focuses our attention on the performers rather than their performance, but that's a matter of tone or shading, rather than meaning properly speaking. Since Davidson's theory accounts for the neatness of the adverb-dropping inference from "They performed well" to "They performed," by assimilating it to the case of "Their performance was good" implying "Their performance occurred," Davidson's theory holds up well here.

Consider related sentences, without adverbs: "Their performance occurred," and "They performed." The first one explicitly quantifies over the performance. The second one could be interpreted to assert that the referent of 'they' satisfies the predicate 'performed.' Here, without adverbs present, the second theory seems better, since Davidson's account lacks point. But of course Davidson's theory would be false if it lacked application here. The point of Davidson's account is that the implication between these sentences of first-order logic:

$\exists$x(x was a performance & x was by them & x was good)

∃x(x was a performance & x was by them)

...can account for the implication between:

They performed well.
They performed.

If we parse the second of those as a simple subject-predicate sentence, say, Pt, then the implication seems to have been lost. So Davidson's theory has to apply in the case where there are no adverbs. Nevertheless, it seems that Pt is a better first-order interpretation of "They performed" than ∃x(x was a performance & x was by them).

On the truth-conditional account of meaning, a sentence's truth-condition is its meaning. We should ask what is the truth-condition of Pt, but also of ∃x(x was a performance & x was by them). Can either of these sentences be true without the other being true? Certainly not; and that doesn't seem to be accounted for with reference to mere necessary equivalence. The very same event, the performance, seems to make both sentences true. In the absence of adverbs, it doesn't matter whether we quantify over events or not; the event makes the sentence true either by being that over which the sentence quantifies, or by being the satisfaction of the predicate by the referent of the subject. The two adverb-free sentences mean the same.

Recall the challenge: if Davidson's theory were true, we would think "Their performance was good," rather than "They performed well." If we can somehow stomach this consequence, then we will have passed the gamut Bennett has laid out for Davidson. But we should accept this consequence, in revised form. If Davidson's theory is true, then it doesn't matter whether we think that their performance was good, or rather that they performed well. The two sentences have the same meaning. When we think the one, we think the other. The only difference between them is a subjective difference in tone or shading.

The other challenge, the one involving causation, is more complicated. It begins with the notion of an event and its *companion fact*. Bennett has names of events, which

take the form S-P-T. 'S' refers to the subject of the event; 'P' refers to a trope, an instance of the property P*, that the subject exhibits, and 'T' refers to the time at which S exhibited P/P*. In "Willow's going crazy yesterday," for instance, Willow is the subject of the event, going crazy is a property that she exhibited (and this event name involves her particular going crazy), and yesterday is the time at which Willow exhibited going crazy.

The companion fact of an event is that fact which we assert to obtain by asserting that the event occurred: for instance, "Willow went crazy yesterday" asserts the obtaining of the companion fact of the event named by the event name "Willow's going crazy yesterday." If 'x' names the event, then 'F(x)' will name its companion fact. (*ibid*. pp. 128-9)

Allowing e_n, for a given value of n, to refer to an event, and C(x,y) to assert that x caused y, Bennett has a proposal for the meaning of causal assertions of the form $C(e_1,x)$: Bennett's view is that $C(e_1,x)$ means that some fact that is part of $F(e_1)$ caused x, whatever x might be. If events were facts, then to say that e_1 caused something would be to say that $F(e_1)$ caused something, but, apparently, it isn't. This quotation needs to be extensive, since Bennett's account is complicated:

> Fact causation rests ultimately on the idea of one fact's being a [necessary part of a sufficient] condition of another. So, if $C(f_1, f_2)$ is true and f_3 is a much stronger fact than f_1, then $C(f_3,f_2)$ will not be true, because f_3 will be much too rich, too strong, for it to be true that without *it* some sufficient condition for f_2 would have fallen short of sufficiency.... If events were facts, then "$C(e_1,e_2)$" would mean that the companion fact of e_1 was a cause of the obtaining of the companion fact of e_2; in the vast majority of cases, however, the former fact is much too rich to be a [necessary part of a sufficient] condition of the obtaining of the other fact. Just yesterday, the job I did in my garden caused a backache. If "the job I did in the garden" named a fact, it would be something like
>
> > the fact that without any preliminary warming up I spent 40 minutes vigorously raking and carrying leaves from a large maple tree, getting them off the lawn and onto the other side of the driveway, using a new plastic rake, alternating between left-sided and right-sided sweeps with the rake...
>
> plus some more. But *that* fact was not an [necessary part of a sufficient] condition my getting a backache, for there was no sufficient condition for the backache that

> needed *that* fact in order to be sufficient; a small part of the fact was all that was needed, namely that I worked vigorously for 40 minutes without a preliminary warm-up. (*ibid*, p. 136)

Bennett is giving a mereological account of fact-identity and fact-causation. A fact can consist of many parts. To have a cause, we need a sufficient condition for the effect. The sufficient condition is a fact. Some parts of that fact are necessary to its sufficiency to bring about certain of its effects; others are incidental to that sufficiency. If a fact C is necessary to some larger fact S's (of which it is a part) being sufficient for the existence of some further fact E, then C is a cause of E. The account does not suggest that S must be a *proper* part of C, so the account includes the case in which C, by being S, is itself sufficient for E, as the special case in which C is not *a* but rather *the* cause of E. (*ibid.*, pp. 44-5. I ignore further specifications Bennett introduces to handle cases of causal overdetermination and transitivity in causal relations.)

The reason that e_1, Bennett's raking his garden, cannot be the cause of his back's aching, is that the companion fact $F(e_1)$ (the fact that Bennett raked his garden) is too rich. The problem is not that $F(e_1)$ is itself sufficient for the effect; it is not. The problem is that only a small part of $F(e_1)$ (the fact that Bennett worked out for 40 minutes without a workout) is necessary for the larger fact (that Bennett did a 40-minute work-out without warm-up, and Bennett's back was such that that sort of work-out would damage the muscles of his back, and such damage is experienced as pain the next day, and...) to be sufficient for the backache. Other parts of $F(e_1)$, such as that the rake with which Bennett was raking was new, are not necessary for the sufficiency of the complete cause.

For Davidson, causation is an extensional relation between events. Explanation, on the other hand, is an intensional relation between descriptions of events. Bennett goes wrong by confusing causation with explanation. Consider the case of Bennett's backache. Bennett's raking his garden with a new rake was the cause of Bennett's backache. Bennett's working-out without stretching was also the case of Bennett's backache, because Bennett's raking his garden with a new rake was identical with Bennett's

working-out without stretching. The fact that raking his garden with a new rake fails to *explain* the backache is a failure on the part of the description we have chosen. The pertinent description is the one that calls our attention to the features of the event that have a lawlike relation (we may suppose) with backaches.

Where Bennett speaks of necessary parts of sufficient conditions, Davidson would speak of descriptions that are necessary, or sufficient, for an explanation of the effect. To explain the backache, it was necessary that the raking not follow stretching. To be sufficient, there were any number of descriptions that had to be satisfied. If we say that the aching of Bennett's back was caused by the raking Bennett did with a plastic rake, we will have said something true, though explanatorily worthless. The description of the cause has no explanatory relationship to the description of the effect, even though the cause described has a causal relationship to the effect described.

The choice here is between Bennett's mereological account of fact-causation, and Davidson's description-oriented account of event-explanation. Where Bennett chops up the causing fact into many parts, each of which is a necessary condition on the sufficiency of the cause's causing the effect, Davidson leaves the cause a unified event, but distinguishes between the explanatory value, not of the parts of the event, but of different descriptions of it. Davidson's account is preferable. Obviously, I can't seriously adjudicate the dispute here, but I can point out considerations that strongly support Davidson's view.

The first consideration is the sheer bizarreness of Bennett's view. What exactly are parts of facts? If we have a conjunctive fact, then we might imagine that the conjunct facts are parts. But the fact that Bennett raked his garden with a plastic rake is not plainly divisible into two facts: the fact that he raked his garden, and the distinct fact that he used a plastic rake. If it weren't for the first "part," the second distinct "part" couldn't obtain. On the other hand, if he hadn't used *some* rake, he couldn't have raked his yard. With this sort of deep interdependence, it's not plain that the notion of a part has any purchase. It's

just not plain what sense we can make of raking the garden with a plastic rake being a part of the raking of the garden. "With a plastic rake" isn't removable from the whole of which it is allegedly a part, so what is the application of the concept of a part?

The second consideration has to do with what Bennett would have to do to articulate the notion of fact-causation (as distinct from allowing that events are facts, and saying something about event-causation). The general notion of a fact, or truth-condition, is one that can't be fully spelled out. But one is sceptical that Bennett could spell out his fact-causation without spelling out facts.

Bennett would have a choice to make between trying to articulate the relations of necessity and sufficiency that inform his account, and claiming that these concepts are baseline. If they are baseline, then by introducing fact-causation, he has made little progress on clarifying the nature of causation. Davidson's account, which makes causation a basic concept, is just as clear and substantially simpler. If they are not baseline, then Bennett must explain how sufficiency can relate facts, without spelling out just what a fact is. Perhaps that is a task that can be performed, but I doubt it.

Bennett's argument against events' being facts was just that, if events were facts, his account of fact-causation would be mistaken. But it is. So there is no particular reason to deny that events are among the facts.

2 Meaning and Interpretation

2.0 INTRODUCTION

In chapter 1, I put forward the view that truth is correspondence to truth-conditions, but that there could be no theory of truth-conditions. By taking a Tarskian theory of truth, along with a "no-theory" theory of truth-conditions, as our theory of meaning, we produce a "semantic theory of meaning" that seems to satisfy three classical and standard conditions on such theories:

> Publicity: meanings can be grasped on the basis of publicly available evidence.
>
> Finitude: infinitely many possible sentence meanings can be grasped on the basis of a grasp of finitely many meanings.
>
> Extensionality: sentence meaning determines truth-condition.

That the theory satisfies Extensionality is obvious. That it satisfies Finitude is a consequence of the structure of Tarskian theories. How exactly it satisfies Publicity is the subject of most of the rest of this chapter.

The first section of this chapter considers two of the main alternative approaches to meaning, by way of considering their objections to the truth-conditions theory. I begin by discussing Dummett, whose objection is precisely that the truth-conditional theory cannot satisfy Publicity. I contend that it can; moreover, Dummett's use-theoretical conception cannot. Then I discuss the Fregean objection (specifically, David Sosa's very clear articulation of it) made standard by "Sense and Reference." I conclude that the Fregean objection shows that attitudes require characterization beyond their content, but that this fact has no implication for theory of meaning. That is an important fact when it comes to moral reasoning.

The following two sections offer the substance of my Davidsonian account of meaning and the attitudes. Davidson picks up Quine's idea of radical translation, and argues that this hermeneutical approach to language and the attitudes supports a number

of subtle and interesting theses. From my point of view, there are two key theses. The first is content externalism, the thesis that the content of assertions and beliefs is external to the mind. Externalism is a commitment of any theory that satisfies Publicity, and the truth-conditional theory of meaning is the leading contender for satisfying Publicity; only the use theory seems poised to satisfy Publicity, and I contended that it could not. The second key thesis is holism. The attitudes, I contend, are holistic because of their intensionality. Here, I contend that the truth-conditional theory can account for the facts Frege offers as an objection to externalistic theories of meaning.

The last two sections consider secondary but still important issues. First, I discuss the indeterminacy of meaning. Semantic indeterminacy is a consequence of satisfying the publicity constraint on theories of meaning. Since any theory must satisfy that constraint, we're stuck with modest indeterminacy. Second, I discuss self-knowledge. What motivates a concern for this issue? Strictly, it's extraneous to my line of argument. However, that externalism can't account for self-knowledge is offered as a main line of criticism to the externalistic anti-sceptical argument I offer in the next chapter. So I concern myself with the paradox of self-knowledge given the hermeneutical approach to, and externalistic nature of, the attitudes. I contend that externalism does not block self-knowledge as long as we have a correct account of what it is to know oneself.

2.1 OBJECTIONS

In section 2.1.1, I consider Michael Dummett's response to the truth-conditional theory of meaning. Dummett makes two objections, the first striking at the heart of the argument for the theory: Dummett denies that the truth-conditional theory satisfies (I). His second objection is that no one could possess a Tarskian theory anyway. In section 2.1.2, I consider the classical Fregean objection that denies that the theory can handle attitude attributions.

2.1.1 Dummett's Objections

Dummett announces the doctrine of meaning as use, but then explains that this is not a theory on a par with meaning-as-truth-condition, or meaning-as-method-of-verification: "The slogan 'Meaning is use' is... of a different character: the 'use' of a sentence is not, in this sense, a *single* feature; the slogan simply restricts the *kind* of feature that may legitimately be appealed to as constituting or determining meaning." (Dummett 1974a, p. 222-3) What is the restriction? "The meaning of... a statement cannot be, or contain as an ingredient, anything which is not manifest in the use made of it, lying solely in the mind of the individual who apprehends that meaning..." (*ibid*, p. 216) As Dummett uses the phrase, "the use theory of meaning" just *is* what I've called the Publicity constraint.

Dummett gives two arguments that amount to critiques of Davidson's view. The first is that the truth-conditional theory of meaning fails to satisfy the Publicity constraint/use theory of meaning:

> On a platonistic interpretation of mathematical theory, the central notion is that of truth: a grasp of the meaning of a sentence belonging to the language of the theory consists in a knowledge of what it is for that sentence to be true. Since, in general, the sentences of the language will not be ones whose truth-value we are capable of effectively deciding, the condition for the truth of such a sentence will be one which we are not, in general, capable of recognising as obtaining whenever it obtains, or of getting ourselves into a position from which we can so recognise it. Nevertheless, on the theory of meaning which underlies platonism, an individual's grasp of the meaning of such a sentence consists in his knowledge of what the condition is which has to obtain for the sentence to be true, even though the condition is one which he cannot, in general, recognise as obtaining when it does obtain. This conception violates the principle that use exhaustively determines meaning.... (*ibid*, 223-4)

Dummett's talk of the platonistic theory of mathematics ought not lead us to believe that he is addressing a narrow, local issue in philosophy of mathematics. What he calls 'platonism' is just the truth-conditional theory of meaning, and this argument against it is fully general and appeals to no special feature of mathematical discourse. This argument strikes at the basis of the support for the truth-conditional account of meaning.

However, Dummett's version of the publicity constraint is not quite right, or at least not as explicit as one would like. It's not use that must exhaustively determine the meaning of an expression, it's all public features of the expression. That includes not just use, but context of use. Perhaps Dummett intends to include context within use, in which case he has merely left implicit something fairly important.

Dummett's claim, then, is that we can't grasp the truth-conditions of an assertion on the basis of publicly available evidence about that assertion. But for anything to be the meaning of an assertion, it must be graspable on the basis of publicly available evidence. Hence, truth-conditions are not meanings. But why can't we grasp the truth-conditions of an assertion on the basis of publicly available evidence? Dummett says, with respect to sentences that can't be proved or disproved:

> Since the sentence is, by hypothesis, effectively undecidable, the condition which must, in general, obtain for it to be true is not one which we are capable of recognising whenever it obtains, or of getting ourselves into a position to do so. Hence any behaviour which displays a capacity for acknowledging the sentence as being true in all cases in which the condition for its truth can be recognised as obtaining will fall short of being a full manifestation of the knowledge of the condition for its truth: it shows only that the condition can be recognised in certain cases, not that we have a grasp of what, in general, it is for that condition to obtain even in those cases when we are incapable of recognising that it does. (*ibid*, 225)

The idea is that, if we can't recognize the truth-conditions of the sentence whenever they occur, then what we can display in using the expression is an only partial grasp of the sentence's truth-conditions. So no more than some instances of the sentence's (no doubt multitudinous) truth-conditions can constitute the meaning of the sentence.

Consider, for example, some sentence Φ. Native speakers/parents/other people seem to assert Φ just under the condition that they become aware of the presence of a rabbit. What we can correlate asserting Φ with, then, is not the presence of rabbits, but just presences of rabbits of which the speaker is aware. But what we really display, then, are not the truth-conditions, but the belief-conditions, of the sentence. Thus meaning is either: some idiosyncratic collection of instances of truth-conditions, or else: belief

conditions. The latter seems more plausible. But this dialectical situation can be turned around. If we can show that meaning is either: some idiosyncratic collection of instances of belief conditions, or else: truth conditions, then the truth-conditional view would be more plausible.

Surely there are at least as many belief conditions for a sentence as there are truth-conditions for it: at least one such condition that puts one in a position to know that the truth-condition holds. But it will be the unusual sentence that has a truth-condition that can only be checked one way. I can see the rabbit out of my left eye, or my right, or I can hear about it, or I can hear it, or I can see it from a foot to the left of where I was standing.... For mathematical statements, there are of course infinitely many proofs (possession of which are belief conditions) for each statement. We don't know most of those proofs, but on the discovery of a new one, we can sometimes recognize it as a condition of belief for a sentence we already understood. We don't thereby experience ourselves as changing the meaning of the sentence by adding a new belief condition; in fact, we don't really think of finding a new proof of an old sentence as anything having to do with meaning at all. Knowledge of truth-conditions can easily outstrip the *available* conditions of belief, which are the only conditions that one can grasp on the basis of publicly available evidence. So we aren't forced to choose between a few instances of truth-conditions or else all of the belief conditions; we're forced to choose between a few instances of truth-conditions or else a few instances of the belief conditions.

And on the other hand, it's not at all difficult to exhibit an understanding of the complete set of truth-conditions for an assertion: anyone who utters a true, lawlike Tarskian T-sentence for an assertion, displays all of the truth-conditions of the assertion. Dummett would respond with the second of the arguments I mentioned above:

> An ability to state the condition for the truth of a sentence [e.g., with a true, lawlike T-sentence] is, in effect, no more than an ability to express the content of the sentence in other words. We accept such a capacity as evidence of a grasp on the meaning of the original sentence on the presumption that the speaker understands the words in which he is stating its truth-condition; but at some point

> it must be possible to break out of the circle: even if it were always possible to find an equivalent, understanding plainly cannot consist in the ability to find a synonymous expression. (*ibid*, p. 224)

This is an attack on the idea that a semantic theory can help to account for linguistic understanding. Of course it's true that rattling off T-sentences does not demonstrate an understanding of a language. But the structure of a Tarskian truth theory nevertheless solves the problem Dummett presents.

Dummett accepts, in the statement of his argument, that we can acquire partial grasp of truth-conditions — grasp of some of the truth-conditions, the ones that are also belief-conditions — on the basis of publicly available evidence. His complaint is that we're limited to this partiality. My task, then, is to show how we can bootstrap from partial to total grasp. To do so, I will appeal to the holistic nature of Tarskian truth theories.

Assume that we're trying to translate/interpret/learn some sentence Φ, which is affirmed by natives/other speakers/elders only under condition that they can confirm that a rabbit is present. For Dummett, Φ means the circumstance that a rabbit is known to be present, whereas for me, Φ means the circumstance that a rabbit is present. But consider Dummett's belief conditions. If the speaker says that Φ under conditions under which there is no rabbit present, but, instead, she thought she saw a rabbit, will we interpret Φ to mean "It seems like there's a rabbit"? If so, we will radically misinterpret an enormous amount of speech. For instance, if such a theory were correct, then no one who learns English will have any idea what the point is of saying, "It seems like," since such a prologue would be implicit in every observation sentence. If not, then we're beginning to be selective about which belief-conditions will count as part of the meaning. If we want to avoid any slippage due to a misinformed informant, we'll have to select exactly those belief-conditions in which the belief that Φ is true.

Φ is asserted under the conditions that a rabbit is believed to be present. How, then, can it mean, not that a rabbit is confirmed to be present, but that a rabbit *is* present?

If we leave Φ at the grammatical level of the complete sentence, thereby severing its connections to the rest of the language of which it is an expression, the meaning will indeed be limited to conditions known by speakers to obtain. But as we begin to learn the grammatical structure of Φ, we put ourselves into a position to grasp its meaning on the basis of the other utterances in which Φ's words appear: that is, we can deduce Φ's meaning on the basis of the axioms of a Tarskian truth theory. Ψ, for instance, is uttered under the conditions that Anya is confirmed to be afraid of rabbits. Φ and Ψ share in common the element α, which we might reasonably suppose is a word used to refer to rabbits, or perhaps it is a predicate satisfied by all and only rabbits. By accumulation of these sentences, we can begin to construct the axiomatic infrastructure of a Tarskian truth theory. Rather than try to determine the meaning of each sentence individually, we rely on the overall structure of the theory to determine the meanings of sentences.

How does this help? The problem was that knowing a T-sentence for a sentence of a language did not qualify one as understanding the sentence. But if one knows T-sentences for *many* sentences of the language, and one knows them on the basis of axioms of reference and satisfaction that are confirmed by large bodies of evidence, then uttering T-sentences *does* count as displaying knowledge of truth-conditions. The confirmation of the T-sentences of a theory confirms the theory as a whole, so all of the experience that has helped to yield the entire theory can be displayed in any given T-sentence that the theory itself yields. Tarskian theories, as we will see in 2.2, can be quite worldly, embedded things.

Dummett said that we must break out of the linguistic circle; but that is not apparently necessary. Since the entire Tarskian theory must be supported by empirical evidence, we enter the linguistic circle only through the gamut of worldly experience. Tarskian truth theories are not circles that imprison us, and we need not be liberated from them.

My solution might seem a touch extravagant. While it's no doubt true that anyone who knows a true, empirically confirmed, Tarskian truth theory for a language, and knows that she knows it, and knows that it's on this basis that she understands that language, understands that language, surely such knowledge doesn't account for anyone's actual knowledge of a language. As Stephen Schiffer says:

> ...whether or not knowledge of the kind alluded to in [a true, lawlike Tarskian truth-theory meeting certain empirical constraints] would, if one had it, suffice for understanding a language, it seems very clear that no actual speaker has such propositional knowledge. (Schiffer 1987, p. 116)

Tyler Burge would agree. He likens linguistic understanding to perceptual judgment, which does not, he says, flow from any sort of reasoning, and is not based on any sort of theory or other judgments about context:

> Linguistic training gives one a reliable understanding of what others say. Status as a competent understander and normal use of associated conceptual apparatus yields a defeasible warrant that obviates the need for evidence or justification. Justification is needed only when anomalies arise, or when one cannot rely on the transformations afforded by presumptive overlap with one's own idiolect. It is not a from-the-beginning open question what someone else says, if a reliable understander presumes on seeming immediate understanding. The understander is prima facie entitled to immediate presumptive understanding. (Burge 1999, p. 242)

Of course it's true that interpretation of another speaker doesn't start from scratch each time one needs to interpret. But the fact that one doesn't start anew each time one interprets doesn't imply that there is no basis whatsoever for interpretive judgments, or that that basis is some sort of *terminally* inchoate skill that can't be accurately captured by an explicit semantic theory.

The analogy with perceptual judgments does not obviously have the weight Burge wants it to have. Is it true that we would regard someone's perceptual judgment as knowledge if she did not possess appropriate contextual judgments (including other perceptual judgments) about the nature of the situation in which the original perceptual judgment was made, that support the validity of that judgment? Consider the most

elementary sort of example: this table is brown. Would we say that I know that the table is brown if I didn't also believe (implicitly) that I'm seeing the table in decent light? Would we acknowledge that I know that I see the table in decent light if I didn't believe (implicitly) that these light bulbs put out normal light, that those windows admit normal light — and that these are light bulbs, those are windows, that this is light.... It seems intuitively plausible to me that when we attribute even the most elementary perceptual beliefs, we must attribute the existence of some range of supporting (perceptual) beliefs about the context to the effect that the situation is one that supports the veridicality of perception even in lieu of extraordinary checks.

Davidson himself accepts the claim that Schiffer and Burge make:

> ...claims about what would constitute a satisfactory theory are not, as I said, claims about the propositional knowledge of an interpreter, nor are they claims about the details of the inner workings of some part of the brain. They are rather claims about what must be said to give a satisfactory description of the interpreter. *We* cannot describe what an interpreter can do except by appeal to a recursive theory of a certain sort. It does not add anything to this thesis to say that if the theory does correctly describe the competence of an interpreter, some mechanism in the interpreter must correspond to the theory. (Davidson 1986, p. 96)

This passage is opaque in intent. In what sense must the theory "correspond" to the competence of the interpreter, if to say that the interpreter has, believes, or knows the theory would be false (since it is a claim about propositions, but not about propositional knowledge of the interpreter)? It's not plain what the significance of a theory would be for interpretation if nobody possessed it: how could we learn, e.g., that interpretation is holistic in nature by studying the holism of a theory that nobody actually uses for interpretation?

The problem with which I'm trying to deal is a challenge by Dummett that uttering T-sentences does not display linguistic comprehension. My reply is that, while that is true, knowing a T-theory within the appropriate context of a semantic theory and on the basis of the appropriate empirical basis does constitute linguistic comprehension. But if Schiffer, Burge, and (alas) Davidson are right, my solution might seem a bad one

because no one could actually display her understanding of a language by stating T-sentences for its sentences, because no one actually knows such a theory. I disagree: I want now to argue that ordinary speakers are in possession of Tarskian truth-theories for the languages that they understand.

Here I will contend that Schiffer and many others are afflicted with a form of philosophical hubris: they think that the vulgar are too simple-minded to grasp something as sophisticated as a Tarskian truth-theory. I disagree. The problem with explaining Tarskian theories to non-philosophers is that ordinary people expect philosophical theories to be subtle and esoteric, not ordinary and commonsensical. Since Tarskian theories are so *obvious* to *everybody*, and since they express knowledge that *everybody* has, the only problem the vulgar have with them is their apparent pointlessness. ("Do philosophers really sit around repeating commonplaces to one another?" they might ask.)

Here is an argument suggested by Ernest LePore.[26] Let us, with Burge, adopt what seems to me to be to be an implausibly extreme form of externalism about perceptual justification: a belief that is true and caused by some procedure that reliably causes only true beliefs counts as knowledge, even in the absence of background (perceptual) beliefs about the context supporting the notion that the first belief was caused appropriately to be veridical. Even if this sort of view applies to perceptual judgment, obviously it doesn't apply everywhere. Imagine that a student runs reliable logic software, and comes to believe every theorem that the software proves. We might not be apt to say that the student knows these theorems, and we certainly would not say that the student knows them if the student doesn't know (much less doesn't believe) that the software is reliable.

Likewise, consider someone who believes, based on a speaker's assertion, "I believe that ϕ," that the speaker believes that ϕ. If the belief that the speaker believes that ϕ were caused by the speaker's utterance in much the same way as the belief that there's a bird is caused by the bird's being their, then we might not be apt to say that the belief

[26] See LePore 1999, esp. pp. 60-1.

constituted knowledge. Beliefs based on other people's autobiographical utterances about their mental states won't count as knowledge without being based on other beliefs with logical relations to the belief in question. Mere reliable habit won't make for knowledge in the case of linguistically based attitude attribution.

To pump the intuition further, let's consider cases. Let's say that you ask me why I believe that there's a bird there, and I say, "Got me. Looks like there's a bird there." We might, with Burge, be inclined to find this adequate to count me as knowing that there's a bird there (if there's a bird there and I wasn't hallucinating or in some Gettier situation). But let's say that you ask a listener why he believes that the speaker believes that ϕ. He replies, "Got me. Sounds like she believes that ϕ." Let's further assume that we think that we're really getting the full story: he genuinely doesn't have any further beliefs about what the speaker meant when she said "I believe that ϕ," beliefs that would be pressed into service to interpret that utterance to show that she believes that ϕ. Then I doubt that we would attribute the speaker knowledge that the speaker believes that ϕ (even if she does &c).

What's the difference between the observation sentence, "There's a bird," and the non-observation sentence, "The speaker believes that ϕ?" What makes the latter not an observation sentence? When attributing to a speaker (in this case, the listener whose utterance is "The speaker believes that ϕ") some belief that can be expressed as an observation sentence, we might not demand of ourselves that we also attribute to the speaker some other beliefs that give her a reason to believe the observation sentence. Perhaps we only have to think of her as having been caused to hold the belief. However, to attribute some belief that can't be expressed as an observation sentence, we're not really *interpreting* the speaker unless we also attribute to her some *reason* for believing the sentence. But, as Davidson says, "...nothing can count as a reason for holding a belief except another belief." (Davidson 1983, p. 141) Interpretation requires some sort of

supporting theory, and that is *obvious* when we move beyond interpreting observation sentences.

While very few speakers have ever articulated it, all of them have tacit semantic knowledge of approximately the sort codified in a Tarskian truth theory. When we try to explain what we mean by an utterance, we tap into that knowledge, and doing so affords us the opportunity to display the truth-conditions of sentences.

Dummett's attack on the truth-conditional theory of meaning was that we could not exhibit a grasp of truth-conditions, only of belief-conditions. I responded that we can: we can exhibit it through saying (or saying things that convey or imply) appropriate Tarskian T-sentences. The fact that such a theory has a holistic structure, and is as a whole based on empirical evidence, means that these linguistic performances are not mere rattlings-off of meaningless synonyms. And the fact that we're apt to be internalists about the justification of content-attributions to other speakers implies that we do have a semantic theory of some kind.[27]

I'm not quite sure what's behind this critique from Dummett, Schiffer, and others. However, I suspect that it's something like this. There's a disjoin between the critics' conception of a Tarskian theory, and what they think is necessary for understanding a language. Understanding a language is, we might say, a warm-blooded behavior. We understand on the basis of long experience and embeddedness within a community of speakers. (To accentuate the mammalian bias here, we might aver to "mother's-knee" learning; a first language has been called one's "milk tongue.") A Tarskian theory, on the other hand, is a mechanical device, cold and formal. The latter could not possibly account for the former: it lacks the richness of real linguistic behavior. The critics see the same deficiency in Tarskian theories that later Wittgenstein saw in the *Tractatus*, and they declare, "Back to the rough ground!"

---

[27] 'Semantic' in the immediately preceding discussion is not intended to include only "semantic" theories in the sense of Tarskian theories, but just some sort of codification of the meanings of utterances.

But why should we assume that Tarskian theories are known in such a cold-blooded fashion? On the contrary, the *general attitude* of Dummett's own use theory of meaning should convince him that knowing a Tarskian theory is a very rich cognitive and behavioral activity. Knowing, on the sort of late Wittgensteinian view that Dummett adopts, is inherently rich and warm and embedded in a form of life. But a Tarskian theory is known on the basis of wide social contacts, broad experience, and subtle distinction-drawing in the context of all different kinds of interpersonal transaction. Knowing a Tarskian theory is as richly connected to the world and social life as it is public, and it is completely public; as public as anything can be. So knowing a Tarskian theory is likewise rich and warm and embedded in a form of life. Such a knowing might well account for linguistic comprehension. Dummett seems to treat the formalism of the Tarskian theory as exhaustive, not just of the known theory, but of the knowing of it; but knowing a Tarskian theory is just what later Wittgenstein thought that knowing is: an embodied expression of a form of life.

2.1.2 A Fregean Objection

In this section, I want to respond to another attack, of a Fregean nature, on the semantic theory of meaning. Obviously the first analytic presentation of this sort of argument was in Frege's "On Sense and Reference." Kripke discusses some versions of this general form of argument under the rubric of a puzzle about belief, each of which centers on (approximately) the substitution of co-referential names; Schiffer offers what amounts to the same argument only centering on natural kind terms. I'll focus my remarks on David Sosa's interpretation and extension of Kripke's discussion.[28]

For Kripke, the point of the discussion of the puzzle about belief is this. A certain Fregean form of argument against the direct reference theory works as a *reductio*, by assuming that the direct reference theory is true and applying it to a group of assumptions

[28] See, respectively, Frege 1892a, Kripke 1979, Schiffer 1987, pp. 55-60, and Sosa 1996.

describing a plausible thought experimental situation. But the same assumptions can generate the same contradiction even in the absence of inferences warranted by the direct reference theory; or, more precisely, it turns out that those inferences can be warranted even without relying on the direct reference theory. Thus, Kripke argues, Frege's *reductio* fails against the direct reference theory: Millianism was not the guilty premise. Sosa argues that even Kripke's reformulated *reductios* tacitly rely on an implication of the direct reference theory. He calls this implication the Hermeneutic Principle, [H]:

> If a name in ordinary language has a single *referent*, then it may correctly be represented logically by a single constant. (Sosa 1996, p. 388)

While there are versions of Kripke's *reductio* that avoid each other possibly relevant commitment — to even the most trivial principles of disquotation or translation — all of them seem to rely on the Hermeneutic Principle. For example, consider version [MA] of the paradox:

(1)	Peter is rational.	Assumption
(2)	Peter, on reflection, assents to 'Paderewski has musical talent.'	Assumption
(3)	Peter, on reflection, assents to 'Paderewski does not have musical talent.'	Assumption
(4)	Peter believes that Paderewski has musical talent.	2, [Disquotation]
(5)	Peter believes that Paderewski does not have musical talent.	3, [Disquotation]
(6)	Peter believes that Paderewski has musical talent and Peter believes that Paderewski does not have musical talent.	4,5, conj.
(7)	If Peter believes that Paderewski has musical talent and Peter believes that Paderewski does not have musical talent, then Peter has contradictory beliefs.	[Hermeneutic Principle]
(8)	Peter has contradictory beliefs.	6, 7, m.p.
(9)	If Peter has contradictory beliefs, then Peter is not rational.	Analytic
(10)	Peter is not rational.	8,9, m.p. (*ibid*, p. 380)

Sosa explains that the Hermeneutic Principle is at work on line (7) on pp. 387-8. The assumptions are obviously unobjectionable. One of several subsidiary principles such as Disquotation is at work in each of Sosa's more explicit displays, but it's not the same

principle in all of them and we can probably generate a version of the paradox that employs none of them (*ibid*, p. 384). Sosa concludes that the problem is [H]. How does [H] help us derive (7)? This argument is not plainly valid:

> [H] If a name in ordinary language has a single *referent*, then it may correctly be represented logically by a single constant.
>
> Therefore,
>
> (7) If Peter believes that Paderewski has musical talent and Peter believes that Paderewski does not have musical talent, then Peter has contradictory beliefs.

Obviously, some further premises are necessary to put [H] in contact with (7). But what should draw our attention is that [H] seems, superficially, an eccentric premise to employ in the derivation of (7). (7) is *obvious*, and should be derivable from uncontroversial general principles, like the definition of 'contradictory beliefs.'

We may characterize contradictory beliefs as beliefs such that the logical representation of one of them is the negation of the logical representation of the other. Such a conception of contradictory beliefs puts the consequent of (7) in contact with the notion of logical representation in [H]. It also opens up enough of a distance between the antecedent and consequent of (7) that we can see a need for subsidiary premises, such as [H]. But why should we open up such a space? Consider the same *reductio*, but allow that 'Paderewski' *is* an ambiguous name: in some instances, it is the name of a famous pianist, in others, that of a philistine politician. If we were to allow (7) under *those* circumstances, then we would find contradictory beliefs that weren't contradictory. So we need some tighter notion of contradictory beliefs. Such a notion is afforded us by the characterization of contradictoriness of belief with reference to the translations of the beliefs into formal logic being such that one is the negation of the other.

In order to deduce (7) from [H], we need additional premises to put the antecedent (A) of (7) in contact with its consequent:

> (A) Peter believes that Paderewski has musical talent and Peter believes that Paderewski does not have musical talent. (assumption for conditional proof)

(H) If a name in ordinary language has a single *referent*, then it may correctly be represented logically by a single constant.

How shall we show the consequent (C) of (7), that Peter has contradictory beliefs? It's plain that we'll need some premise about contradictory beliefs; how about:

(B) $\forall x \forall \Phi \forall \Psi \{[\Phi$ is a representation in formal logic of one of x's beliefs $\wedge \Psi$ is a representation in formal logic of one of x's beliefs $\wedge (\Phi = \sim\Psi)] \rightarrow$ x has contradictory beliefs$\}$

We'll furthermore need something to apply [H] to the name 'Paderewski':

(P) The name 'Paderewski' is a name in ordinary language with a single referent.

The predicate of Peter's beliefs is now a problem (this, incidentally, is where Schiffer's version of the argument makes its intervention, since his puzzle is the same as Kripke's but for its focus on natural kind terms rather than names). Our [H] only governs the translation of names. We'll need an additional Hermeneutical principle for predicates:

[H_p] If a predicate in ordinary language has a single satisfaction condition, then it may correctly be represented logically by a single constant.

And we'll need something to apply [H_p] to the predicate 'has musical talent':

(T) The predicate 'has musical talent' is a predicate in ordinary language with a single satisfaction condition.

With this array of premises available, we may now deduce (7) from first principles:

[H]	If a name in ordinary language has a single *referent*, then it may correctly be represented logically by a single constant.	Hermeneutical Principle
(P)	The name 'Paderewski' is a name in ordinary language with a single referent.	Assumption
(*)	The name 'Paderewski' may correctly be represented logically by a single constant.	[H], (P), $\forall$, $\rightarrow$
[H_p]	If a predicate in ordinary language has a single satisfaction condition, then it may correctly be represented logically by a single constant.	Hermeneutical Principle
(T)	The predicate 'has musical talent' is a predicate in ordinary language with a single satisfaction condition.	Assumption
(**)	The predicate 'has musical talent' may correctly be represented logically by a single constant.	[H_p], (T), $\forall$, $\rightarrow$
(A)	Peter believes that Paderewski has musical talent and Peter believes that Paderewski does not have	Assumption for Conditional Proof

	musical talent.	
(***)	"Tp" is a representation in formal logic of one of Peter's beliefs and "~Tp" is a representation in formal logic of one of Peter's beliefs.	(*), (**), (A)
(B)	∀x∀Φ∀Ψ{[Φ is a representation in formal logic of one of x's beliefs ∧ Ψ is a representation in formal logic of one of x's beliefs ∧ (Φ = ~Ψ)] → x has contradictory beliefs}	Definition of 'Contradictory Beliefs'
(C)	Peter has contradictory beliefs.	(***), (B), ∀E3, →E
(7)	A → C	(A), (C), →

The assumptions define a plausible thought experiment. The definition of contradictory beliefs flows from the necessity that mere paradoxes born of ambiguity not show as contradictions. If (7) gets us into trouble, then, it's because of one or the other of the Hermeneutical principles involved in its derivation, and we're now in a position to see the precise role of those principles.

Lines (1)-(3) of the original argument define a plausible thought experiment, so they're not the problem. (7) is supported by the same assumptions supporting the same plausible thought experiment, plus a plausible definition of contradictory beliefs and the Hermeneutical principles. (9), the claim that if Peter has contradictory beliefs, then Peter is not rational, is held to be analytic. Since (1)-(3) aren't the problem, it must be one of (7) or (9). Since (7) is supported by a group of premises of which the only controversial ones are the Hermeneutic principles, the problem is either the Hermeneutic principles or else (9). But the Hermeneutic principles are supported by the Davidsonian theory that the contribution of a name to the meaning of a sentence is its referent. I'm therefore forced to deny (9): I claim that though Peter has (occurrent, reflective) contradictory beliefs, Peter is nevertheless, for all we know, rational.

Sosa contends that this position "appears to be theory-laden." (*ibid*, p. 391 n.14) He points out that, "...the contradictory beliefs in question are *logically contradictory* (not merely incompatible) — their form would be such that one is simply the negation of the other. No difficult procedures would be employed to ascertain their inconsistency." (*ibid*)

As I've tried to make graphically clear, that last claim is not true. Two Hermeneutical principles are necessary for those of us with a knowledge of formal languages to ascertain the inconsistency, assuming that we use the sort of discerning account of inconsistency embodied in (B).[29] But that account can only be comprehended by those of us with a knowledge of formal languages. If we use a correct account of contradiction, then only speakers of first-order logic who know the Hermeneutical principles could discover certain contradictions. Surely, then, things are not so easy.

However, we may assume, with Kripke, that Peter "is a leading philosopher and logician. He would *never* let contradictory beliefs pass." (Kripke 1979, p. 122) Since all leading philosophers and logicians know formal languages and the correct account of contradictory beliefs, Peter, one might think, can discover contradictions just as well as we can. So Peter can run the argument [MA] that I'm discussing, diagnose the problem (that is *how* he refrains from letting contradictory beliefs pass), and reject some belief or other.

However, to run [MA], Peter needs (7). To get (7), he needs to believe all of the various premises I introduced to get (7) from [H]. Among those premises is one that he doesn't believe, (P): The name 'Paderewski' is a name in ordinary language with a single referent. Peter believes that 'Paderewski' is an ambiguous name in ordinary language, so he cannot derive (7) from [H]. He is therefore not in a position to discover the contradiction that he believes, because he is not in a position to get his contradictory beliefs into explicitly contradictory form.

Line (9), I propose, is false. Believing contradictions is not irrational, because it's possible to believe two sentences, one of which would be the negation of the other if both were translated into a formal language, without knowing that they are contradictory. Line (9) is superficially attractive because it's near to the more plausible (if still approximate) (9'): If Peter has contradictory beliefs and knows that he has contradictory beliefs, then

[29] ...and inspired by Sosa's example of Rock's beliefs about Paris and Paris; *ibid* pp. 386-7

Peter is not rational. (9') wouldn't lead to a contradiction, because Peter does not know that he has contradictory beliefs.

I want to make clear just what I take my commitment to [H] to amount to. [H] says that, if two tokens of one word refer to the same object, then those two tokens may be translated into the same constant of an appropriate formal language. But this is not all to which we should commit ourselves. We should commit ourselves to:

> [H*] If one (*or more*) name(s) in ordinary language has (have) a single *referent*, then it (they) may correctly be represented logically by a single constant.

How shall we apply this? Return to Peter, but give him some beliefs about Hesperus and Phosphorus. He has somehow acquired these beliefs: that Hesperus is hot, and that Phosphorus is not hot. But 'Hesperus' and 'Phosphorus' are one or more names in an ordinary language that have a single referent; thus, they may correctly be represented logically by a single constant. We may, if we wish, represent these as Hh and ~Hh. Peter would not do that, of course, since he doesn't know that the two names share referent. But as we from the outside try to ascertain the consistency of his beliefs, we reveal a problem by making the translation in this way.

Early Wittgenstein committed himself to [H*] when he wrote 5.53: "Identity of the object I express by identity of the sign and not by a sign of identity." But his denial of the significance of identity claims is immediately followed by his rejection, at 5.54, of intensionality. Denying that difference of natural language word could be represented by a difference of formal language translation would bring in train the rejection of intensionality in formal languages. If we *could not* represent 'Hesperus' and 'Phosphorus' by different constants, then we *could not* display, in a formal translation of Peter's beliefs, that Peter is not irrational. But surely we *could* do so, if we wanted to.

Notice that the Hermeneutical principles are all written as *permissions*. Various translations into formal languages might be done for various reasons. If we're trying to capture contradictions, then we're trying to capture beliefs both that a truth-condition

obtains and that it doesn't. To capture contradictions, then, we should follow the procedure of Wittgenstein's 5.53 and translate all co-referential terms with the same constant, since only that translation will avoid appearing to multiply truth-conditions. But capturing contradictions is not the only task we might perform by translating into a formal language. We might be trying to simulate some actual reasoning done by Peter, for instance, or gain insight into his plans or beliefs. In that case, it would be absurd to translate co-referential terms to the same constant (even if they're the same word, as with 'Paderewski'). We should translate more or less as Peter himself would, if we're trying to capture his beliefs as he understands them.

Contradiction is absolute, but the appearance of contradiction is relative to a language. If translations from a natural language into a formal language are done in accordance with the Hermeneutic principles *and* ordinary empirical facts (like that 'Paderewski' isn't ambiguous), then the formal language can display contradiction perfectly; that is why the appearance of contradiction relative to a formal language is, according to my definition of contradiction, identical with contradiction. For Sosa, denying (9) is "theory-laden," but it seems that *accepting* it is much more theory-laden: for to accept (9) is to accept that everyone knows the translation of their beliefs into a formal language done according to a correct application of the Hermeneutical principles and with a complete knowledge of the identity and difference of reference of all of their referring terms. Earlier, I shockingly attributed to all speakers knowledge of a Tarskian truth theory for the languages that they understand. But accepting (9) requires attributing much, much more explicit semantic knowledge to ordinary people than a mere Tarskian theory.

So far, this treatment has been superficial, dealing as it has with language, and never with what we mean by our utterances. The Fregean challenge is deeper than I've been treating it so far. Here, I take it, is the idea. Meanings are what we grasp when we understand language. They are present to the mind, as they are our thoughts. So we have

special cognitive powers and authorities when it comes to meanings. I can't really go far wrong when it comes to my own thoughts. (This way of thinking leads someone to believe that subjective clarity and distinctness is a mark of truth.) But to accept a contradiction is to accept as true two thoughts that can't both be true. Perhaps I can do that if I'm not very reflective or I don't think the two contradictory thoughts at the same time. But to have such thoughts reflectively and occurrently is impossible, unless I just don't care about whether what I take to be true, is true. And that's one way of being irrational. So, as long as I'm rational, I won't reflectively and occurrently believe any contradictions.

The problem here is with the notion of special cognitive powers that serve to protect a rational person against contradiction. What could justify our belief that these powers exist? Dummett says, about Fregean meanings:

> ...a sense cannot have any features not discernible by reflection on or deduction from what is involved in expressing it or in grasping it. Only that belongs to the sense of an expression which is relevant to the determination of the truth-value of a sentence in which it occurs; if we fail to grasp some features of its contribution to the truth-conditions of certain sentences, then we fail fully to grasp its sense, while, on the other hand, any aspect of its meaning that does not bear on the truth-conditions of sentences containing it is no part of its sense. It cannot be, therefore, that the sense has all sorts of other features, not detectable by us.... A thought is transparent in the sense that, if you grasp it, you thereby know everything to be known about it as it is in itself. (Dummett 1981, pp. 50-1)

While the reasons for this conviction are complicated and perhaps obscure, the notion of a meaning, many are convinced, is the notion of something that is given wholly to the mind; that reserves nothing from the mind to which it is given. Hence contradictions are on the surface as everything relating to meaning is on the surface.

Ironically, senses so construed would fail to achieve the theoretical task Frege designed them to perform. Consider the cognitive value of a sentence like "Hesperus is Phosphorus." Since it can't mean just that Venus is itself, nor that the word 'Hesperus' is co-referential with the word 'Phosphorus,' it must mean that the sense of 'Hesperus' and the sense of 'Phosphorus' determine the same object. But if determining objects were all

that the senses did, then of course anyone with the two senses in mind would immediately ascertain that they determined the same object. That that *isn't* what happens every time someone grasps the meanings is precisely what senses were invoked to explain. If senses aren't at least as obscure as what they determine, grasping them would be sufficient to grasp identities of referents, including truth-conditions. But grasping two senses that are both true under the same circumstances is not sufficient for grasping the identity of the two senses' truth-conditions. So senses must be as obscure as what they determine.

Davidson likewise disagrees with Dummett's view, and the view implicit in Sosa's claim about contradictions and rationality, line (9). As against that Cartesian fantasy of mental contents totally exposed to a mental eye, he writes:

> ...if a thought is constituted the thought it is by the mind's knowledge of an identifying object, then someone knows what thought she is thinking only if she knows which object she has in mind. Yet there seems to be no clear meaning of the idea of knowing which object one has in mind. The trouble is that ignorance of even one property of an object can, under appropriate circumstances, count as not knowing which object it is. This is the reason philosophers who have wanted to found knowledge on infallible identification of objects have sought objects that, like Hume's impressions and ideas, 'Are what they seem and seem what they are' — that is, that have all and only the properties we think they have. Alas, there are no such objects. (Davidson 1989, p. 54)

Fregean opposition to the semantic theory of meaning rests on the assumption that (reflectively believed, occurrent) contradictions are transparent to the mind, and that's believed because the contradictory beliefs are held to be transparent to the mind. Since truth-conditions aren't transparent to the mind — as shown by the fact that contradictions between them aren't transparent — they must not be meanings. But nothing is transparent, so truth-conditions' opacity doesn't disqualify them from serving as meanings.

2.2 Interpretation and Charity

If the presentation so far has been reasonably compelling, theories of meaning other than the truth-conditional theory will seem problematic or unmotivated; at least the

truth-conditional theory of meaning will be seen to have survived some sharp challenges. But it remains to be seen that the truth-conditional theory can satisfy the publicity constraint. To check that, we should consider the thought experiment of radical interpretation.

In radical interpretation, someone who does not understand a speaker must develop an interpretation of the speaker's beliefs and utterances on the basis of no prior knowledge of the language or intentional states of that speaker. The situation of radical interpretation, then, is a test case for our ability to generate Tarskian theories of truth within the publicity constraint. Davidson characterizes the situation like this:

> The evidence cannot consist in detailed descriptions of the speaker's beliefs and intentions, since attributions of attitudes, at least where subtlety is required, demand a theory that must rest on much the same evidence as interpretation. The interdependence of belief and meaning is evident in this way: a speaker holds a sentence to be true because of what the sentence (in his language) means, and because of what he believes. Knowing that he holds the sentence to be true, and knowing the meaning, we can infer his belief; given enough information about his beliefs, we could perhaps infer the meaning. But radical interpretation should rest on evidence that does not assume knowledge of meaning or detailed knowledge of beliefs. (Davidson 1973b, pp. 134-5)

Why does a speaker say, assertively, ϕ? Presumably, because she believes that ϕ means what it does — say, ψ — and she believes that. (Also, that it made sense to say ϕ under the circumstances: it was pertinent.) If we knew what the speaker meant, we could infer what she believed; if we knew of what belief she was trying to assert the content, we could infer what she meant. But we know neither. We have, as it were, one equation with two variables.

Davidson's solution to this problem is to hold belief steady: we assume that we know the speaker's beliefs in advance, so we have only to solve for meaning. But how do we identity the beliefs? The various answers to this question have all been called "the" principle of charity. These proposals all center around the idea that the beliefs of the speaker are approximately what ours would be, were we to find ourselves in the situation of the speaker. Sometimes this notion is put with reference to the idea of finding that the

speaker's beliefs are mostly true, but, since we're the ones making the assignments of truth, finding that the speaker believes (what we believe to be) the truth and finding that the speaker believes what we believe (to be the truth) will be effectively the same assumption. Here's an early statement of the principle: "We want a theory that satisfies the formal constraints on a theory of truth, and that maximizes agreement, in the sense of making [speakers] right, as far as we can tell, as often as possible." (*ibid*, p. 136)

Various of the principles of charity have been subject to many criticisms, of which I'll discuss one below; the rest will turn out to be unhelpful given the results of that discussion.[30] It's reasonably plain that Davidson's own best version of the principle is the principle of optimizing truth: we take the speaker to be saying something true whenever we can. I turn to the discussion by Lepore and Ludwig, which adds considerable clarity to the issue.

The point of the radical interpretation thought experiment is to show how we can go from publicly available evidence to an interpretive truth-theory for a speaker. We can understand this as the task of explaining how we can move from behavioral evidence like (L) to T-sentences like (TF):

> (L) For all speakers *S*, times *t*, *ceteris paribus*, *S* holds *s* true at *t* iff *p*.
>
> (TF) For all speakers *S*, times *t*, *s* is true for *S* at *t* iff *p*.[31] (Lepore and Ludwig 2005, pp. 183-4)

We must somehow move from observing speakers holding a sentence true under circumstances, to those sentences being true under those circumstances. Further, our

---

[30] Beyond the usual suspects, see Cutrofello, 1999, Goldberg, 2004, Vahid 2001, and especially McGinn 1977; for explorations of similar ideas, see Lewis 1974 and Grandy 1973; for basically clueless rebuttals to the principle of charity, see Norris, 1985 pp. 193-217, and Fuller, 1988, pp. esp. pp. 139-62. One sometimes finds Hacking 1979 mentioned as a *locus classicus* contra the principle of charity. Since the alleged criticism allegedly finds its basis in Foucault, of whom I think quite highly, I would be quite concerned about any such criticism. However, I don't find one. The overwhelmingly best treatment of the principle of charity — most sensitive and sympathetic in its interpretation and most trenchant in its criticism — is Lepore and Ludwig 2005, pp. 174-208.

[31] It should (but won't) go without saying that being "true for *S*" is not a matter of truth being relative in the sense of subjective. A sentence is true for one speaker and not for another if there is a difference in the values of indexical expressions appearing in the sentence. "I am Spartacus" is true "for" Spartacus, false "for" the rest of us.

(TF)-style sentences can't be just any materially adequate T-sentences. *p* must give the meaning of *s*; not just some condition that happens to hold iff *s* is true, but the truth-condition of *s*. Ignoring that constraint, the obvious principle to warrant the inference would be:

> (Veracity): For all speakers *S*, times *t*, sentences *s*, *ceteris paribus*: *S* holds *s* true at *t* iff *s* is true (*S, t*). (*ibid*, p. 186)

However, this principle is inadequate. Not just any Tarskian theory that maps true object-language sentences on to true metalanguage sentences will be an adequate theory of meaning: this principle will warrant a Tarskian theory that gives materially equivalent metalanguage sentences for each object-language sentence, but mere material equivalence is obviously not sufficient for interpretation.

Before moving on to a better version of the principle of charity, I want to attend to just one argument against Davidson's principle of Veracity, which is not only a powerful argument in its own right, but also points us in the right direction. Colin McGinn argues:

> ...the motivation for Davidson's principle has much in common with certain assumptions that (in part) prompt a description theory of names: viz, that denotation is fixed by a certain sort of semantic fit between an object and the predicates a speaker associates with a name and supposes *true of* its bearer. And what Kripke-type counterexamples to that theory show... is precisely that reference is autonomous with respect to truth. It is a consequence of this autonomy that a scheme of reference for a given (natural) language cannot be adequately characterized as that total assignment of objects to singular terms which induces a certain distribution of truth values upon those sentences of the language to which its speakers are disposed to assent: e.g., the assignment that *maximizes* truth. If charity recommended such an assignment as determining a scheme of reference, as it does, it would very probably deliver an incorrect scheme. (McGinn 1977, p. 527)

McGinn's complaint about Veracity is that, by assuming that the referent of a word is whatever would maximize the number of sentences in which that word appears that are true, Davidson makes his theory of reference susceptible to Kripke cases. To avoid Kripke cases, Kripke invoked the causal theory of naming; similarly, Putnam invoked the causal theory of reference to natural kinds. But for Davidson, words are theoretical

entities: names and natural kinds terms get their referents because of the sentences in which they appear. We may have that view and also avoid Kripke cases if we relate entire sentences causally to entire situations, rather than relate names to objects. What's necessary both to have an effective principle to take us from (L) to (TF) and to solve the problem McGinn points out is what Lepore and Ludwig call Grace:

> (Grace) *Ceteris paribus*, when we replace '*p*' in (*S*)
>
> (*S*) *S* believes at *t* that *p*
>
> with a sentence that expresses the content of an environmentally prompted belief of *S*'s, the sentence expresses also a condition in *S*'s environment that prompts that belief. (*ibid*, p. 194)

The leading idea is that we should identify the content of a sentence with its cause. Vahid makes what amounts to the same suggestion in different jargon; he suggests that the best version of the principle of charity is:

> (CHc) Beliefs supervene on their causes. (Vahid 2001, p. 317)

Davidson would endorse these suggestions: in a passage to which I recur, he contends that "...we must, in the plainest and methodologically most basic cases, take the objects of a belief to be the causes of that belief."[32] (Davidson 1983, p. 151) A lot of work has to be done here.[33] The first problem is that causes are concrete individual

---

[32] Given the context of this remark and the role played by what we must do, I would suggest that this is another of Davidson's own versions of the principle of charity, though it is not recognized as such by the commentators.

[33] The fact that Grace, not Veracity, is the requisite Principle of Charity provided the dialectical pressure for my complaints about Davidson's own truth-conditions-free account of truth in section 1.3.1-3. If radical interpretation is to proceed only if some contents are causes, then it's important that at least some contents (the ones that are causes) be real, concrete entities, not abstract, theoretical ones, since, only realia are causes. Likewise, it was important in 1.3.4 to establish (however weakly) that events, which are in Davidson's theory of causality causes and effects, could be truth-conditions; else they could not be contents, and either Grace or Davidson's theory of causation would be false. (I'm prepared to entertain that Davidson's theory of causation is false, but I'm no prepared to discuss that tangential matter here.)

events, while contents are no more individual than the general laws by which they are assigned to sentences. So what can it mean that "cause is content"? The second problem is identifying a unique (enough) cause, out of the enormous array of events that occasioned the belief, to be "the" cause that gets to be content (however it is that cause is content).

To solve the first problem, I want to begin by defining two very simple technical notions: the belief set, and the truth-conditional predicate. A belief set is a set of beliefs unified around a common truth-condition: two beliefs belong to the same belief set just in case they have the same truth-condition. This is intended to be a fairly strong condition. Neither material nor necessary equivalence is sufficient for sharing truth-condition. In fact, it's impossible to say what would be sufficient for sharing a truth-condition, since trying to articulate truth-conditional identity would require the prior articulation of the notion of a truth-condition, and this, as I have argued, cannot be done. But the intuitive idea is clear: two beliefs share truth-condition just in case they are necessarily equivalent, and, whether true or false, are true or false for exactly the same reason. It's a consequence of the opacity of meaning that I may have multiple beliefs within the same belief set, for I may have multiple beliefs with the same truth-condition and yet not realize it; and of course, you and I have many beliefs within the same belief set, for you and I have many beliefs with the same content.

The second notion is the truth-conditional predicate. A truth-conditional predicate is some predicate satisfied by each of the members of a set of truth-conditions and nomically correlated with the existence of members of some belief set. There is a defeasible, lawlike relation between the existence of a satisfier of the truth-conditional predicate (in the appropriate context) and the existence of members of a certain belief set. The idea is that the satisfiers of the truth-conditional predicate are the causes of the beliefs. That means, of course, that not all belief sets have truth-conditional predicates: only observation beliefs, beliefs that are caused pretty directly by their truth-conditions,

are members of belief sets with truth-conditional predicates. The satisfiers of the truth-conditional predicate are a set of truth-conditions that, collectively, are the content of the belief set.

This reference to a *collection* is what's to solve the first problem with the simple and rosy formula that "cause is content." The problem with the slogan is that it elides a type-token distinction[34]. Consider my belief, held right now, that it's Wednesday. Consider, likewise, my belief, held a week ago, that it was Wednesday. In one sense, these two beliefs share content; obviously, any utterance of mine of "It's Wednesday" will be true iff it's Wednesday at the time of utterance. But the beliefs also have different contents: the truth-condition of my belief last Wednesday was that it was Wednesday *then*, which is different from its being Wednesday *now*. The sense in which they have different contents is the sense in which they have different *token* causes; the sense in which they share contents is the sense in which they share a *type* of truth-condition.

There is a belief set for all of my beliefs that it's Wednesday (held whenever). There is also a truth-conditional predicate holding of all of the truth-conditions for all of those beliefs. When I say that x's cause is its content, I mean that x's *token* cause is a member of the set of satisfiers of the truth-conditional predicate of the belief set of which x is a member. The cause is a token member of the content-type.

I want to move on now to the second problem, the problem that saying anything about "the" cause of anything is problematic. To speak of "the" cause of anything, including a belief, is to select a point in a long chain of causes. On what basis can we select some one point of this chain that culminates in a belief, and say that *that* is the cause (and satisfier of the truth-conditional predicate)? Here, Davidson introduces the notion of triangulation.

Try to imagine a thinker with no communicative relations to other thinkers. The thinker's experience of the world is, as it were, a point along a chain of causes. The next

[34] ...known sometimes as the character-content distinction.

point out in the chain occurs somewhere in the nervous system; the next point happens at the interface between the sense-organs and the world; the next point happens out in the world where objects and events send energy to strike that interface. Why should we choose one of these points, rather than another, and call it the cause and the content?

In a confused attempt to naturalize epistemology, Quine chose the very last point before internal mental processing begins:

> ...a stimulation σ belongs to the affirmative stimulus meaning of a sentence *S* for a given speaker if and only if there is a stimulation σ' such that if the speaker were given σ', then were asked *S*, then were given σ, and then were asked *S* again, he would dissent the first time and assent the second....
>
> The stimulations to be gathered into the stimulus meaning of a sentence have for vividness been thought of thus far as visual, unlike the queries that follow them. Actually, of course, we should bring the other senses in on a par with vision, identifying stimulations not just with ocular radiation patterns but with these and the various barrages of other senses, separately, and in all synchronous combinations.... (Quine 1960, pp. 32-3)

What belongs to the meanings, for Quine, are stimulations, and stimulations are irradiations of the senses. This is the nearest one can come to sense-data without actually introducing sense-data. The problem with meanings being sense-data, or irradiations of the senses, is that both views make meaning private. Sense-data are, of course, terminally private. And, as some languages are actually known, but nobody knows much about the irradiations of anyone's sense organs (their own or others'), grasping surface irradiations can't possibly be necessary to grasping meanings.

Quine's approach has the advantage of not being arbitrary. The last available point on the causal chain culminating in the belief has a claim on being "the" cause of the belief that it's not obvious that any other point could have. Why should any other point get priority? Davidson's answer to this question introduces a second person and makes thought exist only against the background of communication. For Davidson, only an interpreter's correlating my utterances to *what seems to the interpreter* to be the

prompting circumstance can settle the question which point in the causal chain is "the" cause, and content, of the utterance:

> All creatures classify objects and aspects of the world in the sense that they treat some stimuli as more alike than others. The criterion of such classifying activity is similarity of response. Evolution and subsequent learning no doubt explain these patterns of behavior. But from what point of view can these be called patterns? The criterion on the basis of which a creature can be said to be treating stimuli as similar, as belonging to a class, is the similarity of the creature's responses to those stimuli; but what is the criterion of similarity of responses? *This* criterion cannot be derived from the creature's responses; it can only come from the responses of an observer to the responses of the creature. And it is only when an observer consciously correlates the responses of another creature with objects and events of the observer's world that there is any basis for saying the creature is responding to those objects or events rather than any other objects or events.
>
> ...until the triangle is completed connecting two creatures, and each creature with common features of the world, there can be no answer to the question whether a creature, discriminating between stimuli at the sensory surfaces or somewhere further out, or further in. Without this sharing of reactions to common stimuli, thought and speech would have no particular content — that is, no content at all. It takes two points of view to give a location to the cause of a thought, and thus to define its content. (Davidson 1991, pp. 212-3)

This passage is complicated. It begins with an argument that I think is misleading. The idea behind the opening argument is that nothing counts as similar except when judged similar by an observer. But if I try to judge two things to be similar, my judgments themselves count as similar only when judged similar by an observer. So any two of my judgments count as similar only when judged similar by an observer. But if no two of my judgments count as similar to one another, then there are no relations of synonymy or inference between my judgments, so they aren't really judgments. Hence, to have beliefs, one must be interpreted be an outside observer. This argument suffers from two defects: nominalism, and false alternative. The first premise is that nothing counts as similar except when judged to be similar. Only nominalism would justify that claim, and nowhere does Davidson present an argument for nominalism. He presents an argument against the usefulness of universals for theories of facts, but not an argument against universals. But even if we were to accept the nominalism (and I am inclined to), the

argument takes for granted that if my responses aren't judged to be similar to one another by some *external* observer, they aren't judged to be similar to one another. But why can't I check my own responses, and judge them similar to one another?

As I say, I think that the passage opens misleadingly; nevertheless, it moves into deep waters and navigates them well. Here's the real idea. Two events of type Ψ occur, and each time one of them occurs, I give response Φ. But, beyond both being events of type Ψ, the events are both connected to the responses Φ by similar causal chains. What makes it the case that I'm responding to the events of type Ψ, rather than to some event between the Ψ event and the Φ response, or to the causes of the Ψ events, is that I'm interpreted as responding to the Ψ events. From the inside, I can't tell the difference between the Ψ events and their causes, or between the Ψ events and the effects mediating between them and my Φ responses. From one end of the line, no two points along the line can be distinguished. But if someone from the outside correlates my responses to the Ψ events, and the correlation holds up in a lawlike way, then a standard has been imposed to determine what point along the causal chain culminating in the belief should count as "the" cause of the belief: it's the point counted that way by an observer.

What is the basis for the observer's choice? No doubt the same as the basis for my own similar responses: the Ψ events seem similar and salient. The judgment of salience is the choice of point along the line, and that is the judgment that I can't make from one end of the line.

The role of the interpreter, then, is crucial. Without being interpreted, I literally don't have beliefs, because there is no basis for saying that any of my responses to the world have any particular contents as opposed to a wide variety of causally related contents, and a state with no particular contents is not a belief.

The obvious thing to do is ask about non-observation sentences. Observers are able to determine that they are assertions, and determine their content; yet theoretical sentences are not correlated with well-organized sets of similar causes. Why must *any*

belief be so correlated? Why couldn't there be a thinker without observation sentences? Non-observation beliefs acquire their content from the observation beliefs to which they are connected by relations of inference. In the absence of *any* observation beliefs, we would not have non-observational beliefs. Thus, we must have observational beliefs to have any beliefs at all, and such beliefs must be correlated to sets of causes.

That isn't to say that each and every observation belief is caused by a satisfier of the appropriate truth-conditional predicate. Some, typically false, observation beliefs, are brought about by deviant causes. But if there were no standard sort of cause from which to deviate, the beliefs couldn't be assigned a content and hence wouldn't actually be beliefs. Before one counts as a believer, one must be observed to respond similarly to similar situations.

Couldn't the observations be mistaken? What if a third party were to mistakenly correlate a set of responses to events that had no causal relations to them? An interpreter cannot make up content from whole cloth. The role of the interpreter is to determine which point along the causal chain culminating in a response counts as the content of the response. The causal chain is a pre-existing and absolutely objective reality. An interpreter that assigned causes not along the chain would be mistaken.[35]

What I've been doing here is trying to clarify and defend the principle of Grace, the principle to which I will allude with the slogan "Cause is content." The argument for Grace is that nothing else but an interpreter's assignment of causes to observation beliefs could possibly lend distinct content to any belief, so if cause is not content, there is no content and hence no beliefs.

Recall the role that Grace is to play. It's not immediately obvious that the truth-conditional theory of meaning satisfies the publicity constraint. So we introduce Grace as a principle that licenses the inference from behavioral evidence to a Tarskian theory. The

---

[35] But couldn't the observer assign the wrong point along the line? I take up this issue in 2.4, since it raises an instance of the indeterminacy of content.

question was whether the truth-conditional theory *could* satisfy the publicity constraint: the answer is plainly *yes*.

By discussing Grace along with triangulation, I offered Davidson's argument that content *doesn't even exist* without its causal connections to the world. That is a profound argument and it gets to the deepest depths of Davidson's causal and social externalism. However, the principle of Grace is crucial, so I want to continue to offer reasons for accepting it, even if those reasons aren't as deep. Rather than focusing on what we could call the *metaphysical determination of content*, whatever it is that *makes* such-and-such be the content of a belief, the rest of this discussion will focus on the *epistemic determination of content*, whatever it is that makes such-and-such *seem* to be the content of a belief to someone trying to figure out what the belief's content it. Since we determine content through interpretation, and the paradigm of interpretation is radical interpretation, I return to that situation.

The procedure of radical interpretation is approximately this. We observe the elder/foreigner. We formulate defeasible *ceteris paribus* laws stating what situations seem to prompt her utterances. On the basis of these laws and the principle of Grace, we infer what the contents of those utterances seem to be. On the basis of these inferences, which are the first theorems of our Tarskian theory for the speaker, we begin to formulate axioms from which these theorems are derivable. On the basis of our axioms, we formulate more theorems, which will be testable consequences of our nascent theory. We then proceed as scientists, often revising the axioms of our theory, sometimes rejecting an observation (on the basis of a presumed defect in the usual causal path that should prompt utterances of the observed sort), until we reach an equilibrium point at which our theorems predict a speaker's utterances except in cases in which it seems natural to attribute error (defeat of the relevant causal laws) to her.

I want to urge two (additional, superficial) arguments in favor of Grace. One is a sort of quasi-transcendental argument. That argument will be comparative, and so can

only really be carried off through an exhaustive survey of the objects of comparison. The second argument will be somewhat more definitive, or at least won't suffer from the open-ended, dialectical nature of the first.[36]

Here is the quasi-transcendental argument. The simplest, neatest, least *ad hoc* theory of meaning that can't be refuted is the (strictly, "a") true theory. The truth-conditional theory of meaning is probably that theory, so it is probably the true one. But if the truth-conditional theory of meaning is true, then Grace must be true: for Grace is a precondition on the formation of interpretive Tarskian truth theories in the situation of radical interpretation, which formation is possible iff the truth-conditional theory of meaning is true. Thus Grace is probably true.

As befits its comparative nature, the argument is modest in its conclusion. To eliminate "probably," I would have to compare the truth-conditional theory with every alternative. This I cannot do. However, I have tried to discuss certain plausible competitors with an eye to making them seem less so. All internalist and Platonist theories, it seems to me, fail to satisfy the publicity constraint. The use theory of meaning might satisfy the publicity constraint, but it fails to satisfy the extensionality constraint. The truth-conditional theory meets all three.

Because of its modesty, this argument is disappointing. One wants something more definite. The second argument that I will propose will be in the tradition of Turing: not the Turing of the famous Test, but the code-breaking Turing. (The similarity between the two is really quite striking.)

Let us assume that we know that some object is subject to interpretation. We know that because it is a human being, and it operates a radio, and it operates that radio in an army, and we're fighting that army. We are able to track this particular radio operator. For instance, perhaps he is using WWII-era signals equipment and our skilled signals

[36] I note that the only arguments for Grace discussed by Ludwig and Lepore (see their (2005), pp. 200-8) are transcendental arguments of roughly the sort I'm discussing.

operators can identify the "hand" of the operator, the pattern or beat of his particular signaling. We pay attention to the signals that the operator is sending (to his headquarters? — let's not presume), but, sadly, the operator is signaling in a unique language, spoken only by enemy radio operators, and intentionally formulated to balk attempts to interpret it. Such a language is known as a code, of course; call this one Code. But we can imagine that the situation is somewhat unlike those with real WWII-era codes.[37] Real codes often obey compositionality: just like natural languages, they have finitely many words but infinitely many possible sentences. The enemy's Code, we may imagine, has finitely many possible sentences. Each sentence is communicated with a single symbol, and the operators have a Codebook that amounts to a Tarskian truth-theory without the axioms, in which Code is the object language and their natural language is the metalanguage. Because of this failure of compositionality, we cannot, as Turing did in the actual war, relate individual symbols of Code to individual symbols of German and just crunch possible relations until we had a match that would give us a Tarskian theory, in German, of Code.

Nevertheless, despite this cryptographic difficulty, there is a procedure that we could employ to interpret the operator. We could observe circumstances as they impinge on the operator, and correlate his utterances in Code to the circumstances of the utterances. The procedure would be difficult. On a given day, the operator transmits "𝔖." What prompted the utterance? The sun was shining; the artillery was landing; the supplies were interrupted on their way to the front.... Any of these might have prompted the utterance. But that *that* makes interpretation difficult suggests that discovering what prompted the utterance is in fact key to interpretation. Isn't it *prima facie* plausible that to find the prompting circumstance is to find the meaning?

Assume that, contrary to actual practice, the codebook is retained across time (and not replaced every 12 or 24 hours, depending on the paranoia of the enemy). Then we

[37] For reality, see the classic account in Lewin, 1978, esp. pp. 25-138.

have access to many utterances of "ﻮ." Slowly, by checking to see what might have prompted the utterance, we might begin to develop hypotheses about its meaning. For instance, if "ﻮ" is uttered each and every time we attack the operator's position with tanks, but never when we attack it with artillery or the sun is out, then we might begin to hypothesize that "ﻮ" is true in Code iff the speaker is being attacked by tanks. We could probably rule out that "ﻮ" is true in Code iff it's sunny where the speaker is, or artillery is attacking the speaker.

Of course, we could make room for deviant causal paths; the law connecting "ﻮ" utterances to attacks by tanks would be defeasible. Let's say that we attack them with infantry, and also some fake tanks (big shaped balloons that infantrymen are carrying, or something). The enemy radio operator signals "ﻮ." Of course we do not revise our theory so that our T-sentence for "ﻮ" is now "'ﻮ' is true iff the speaker is being attacked by tanks, or else by fake tanks being carried about by infantrymen." We understand how the causal relations between the operator's utterances and the world have been defeated, so we discount the case. Meaning is *typical* cause. Meanings are assigned to sentence-types, and only through those to their tokens.

We have entered into this situation with the assumption that the enemy radio operator is apt to be saying things of military significance. We discount at the outset the possibility that he is in fact telling his hearer how he is indexing a collection of old china, and that, by some cosmic coincidence, every time he tells his hearer "ﻮ," that is, that he is putting the brightest remaining piece of old china in the leftmost available of the topmost available positions in the display case, we happen to be attacking his position with tanks. We assume that the operator shares a certain set of interests with us, and that he will find salient the same situations in his environment that we do. The more his interests and ours diverge, the harder it will be to interpret him; not because of cosmic coincidences, but because it will just be hard to figure out what situations prompt his utterances. The fact that *that* would make interpretation harder suggests that discovering

utterance-prompting circumstances is crucial to interpretation. Again, that lends support to Grace.

This case is a much better example of interpretation than an example that has been used as the basis for a critique of Davidson's view, the decipherment of Linear B. Wallace argues, on the basis of a discussion of the history of that decipherment project, that Davidson (and Quine) is (are) wrong about how radical interpretation happens.[38] Vermazen has pointed out[39] that the decipherment of Linear B was a decipherment, not an interpretation: the idea was that the Linear B scripts were written in a language already better-understood, but in an unusual character set. Relating one character set to another is not particularly similar to radical interpretation.[40] However, some aspects of the task do seem to be interpretive. Better to point out that we have no reason to believe that the inscriptions in Linear B were "the plainest and methodologically most basic cases" of interpretation, so the method of interpretation will have to be enriched in ways to be discussed in the next section. The principle of charity, however construed, will have an indirect or sophisticated application to this case. Grace is intended to apply directly to observation sentences, only indirectly to non-observation sentences. But as Wallace notes, in the attempt to do the decipherment,

> In order to get on with their task, the interpreters every now and then have to put aside the puzzling inscriptions, and go off and study some aspect of the real world. How many sheep are there in Crete? What is the point of keeping wethers? ...These are some of the questions Killen finds it necessary to take up in the course of his article [on the decipherment]. The interpreter knows what aspect of the real world is relevant, because he has a good idea of the scheme of activity into which the records fit — contextual frame again. (Wallace 1986, p. 230)

By 'contextual frame,' I imagine Wallace to mean 'context, as understood by the speaker (or in this case, scribe).' Why is real-world knowledge necessary for the decipherment? Insofar as the decipherment is interpretive, it would help the interpreters if they knew

---

[38] Wallace 1986, esp. see the summary on pp. 230-1.

[39] Vermazen 1986.

[40] It is, however, more similar to what Turing did in the actual war than what happened in my cryptographic fantasy.

what range of contents they could plausibly attribute to the scribes. But what they seem to be checking for, albeit indirectly, is what sorts of *causes* would be apt to prompt the scribes' utterances (inscriptions). So far from a counterexample, Wallace has produced a weird confirmation of the prediction that Grace will be relied on in cases of interpretation.

2.3 SEMANTIC AND ATTITUDINAL HOLISM

Before I begin in earnest, I want to put the two kinds of holism in touch with one another. I want to address both semantic and attitudinal holism in this section, and I don't want to be too discerning about which one I'm talking about most of the time. Why is this lax procedure permissible?

Recall the problem of radical interpretation. In radical interpretation, we need to solve for both meanings and attitudes. If we knew what someone meant, we could infer her attitudes; if we knew her attitudes, we could discover a semantic theory of truth for her utterances. But we can't discover either one without knowing the other. Because attitude attribution of the peculiarly fine-grained sort we're used to only emerges as a product of interpreting speech, any holism that inheres in the interpretation of utterances will inhere also in the interpretation of the attitudes posited to account for those utterances. If we were to distinguish between two moments of radical interpretation according to whether the focus at that moment is on speech or attitudes, then, since the evidence base for the one interpretive moment lies largely in the other, the holism of the one is automatically the holism of the other. So, for the purposes of this section, I'll routinely ignore *which* holism I'm discussing and allow the arguments to range over both moments of radical interpretation.

In this section, I want to raise three objections to the semantic theory of meaning as so far presented, and solve them by introducing the holistic component of Davidson's view. I'll also address a basic objection against holistic views, and conclude by discussing an apparent inconsistency between externalism and holism.

The first objection is that the account so far doesn't distinguish between rational agents with intentionality, and every other object that can be affected by its environment — i.e., every other object. Consider, for instance, the beach. As I walk along it, I leave footprints. The theory so far has it that an utterance has its cause as its content. But the footprints are effects. Why should I not interpret each of them as an utterance by the beach, meaning, "I was stepped on here"? What's special about effects that are intentional utterances, to distinguish them from all other effects?

The second objection is that the account so far completely fails to pay any attention whatsoever to the intensionality of attitude attributions or semantic content. Whenever water's being wet causes an utterance, H_2O's being liquid at room temperature causes that same utterance. Yet an utterance of "water is wet" is not in every way equivalent to an utterance of "H_2O is liquid at room temperature." So something other than external cause must differentiate utterances and attitudes from other utterances and attitudes with the same cause or truth-condition. But if external cause *is* content, then that condition could not be met.

The third objection has to do with what happens when we move beyond the most methodologically basic cases of interpretation. Perhaps it's plausible to say that, if its' raining around here routinely causes me to say, "It's raining around here," then its' raining is the content of my utterance. But when I say something like, "Neil Gaiman is able to blend many world mythologies into a natural-seeming milieu because of his commitment to a Jungian conception of the existence of narratives," it had better not be the case that in order for my utterance to be interpreted, a would-be interpreter needs to correlate my utterances of this sentence to events of Neil Gaiman's conceptions determining his abilities. Since I've never been in the presence of any such events, this utterance of mine isn't going to have any content, if content is cause. (Perhaps the content to assign it is, "I've been reading a lot of Neil Gaiman," since having read a lot of Neil Gaiman is what triggers the utterance.)

The answer to all of these objections lies in the holistic nature of Tarskian theories and attitude attributions. I have, in a sense, already appealed to this aspect of Davidson's view. As I explained in 2.1.1, Dummett argues that knowing a T-sentence does not qualify one as understanding the left-hand sentence of the T-sentence. That is the complaint that the semantic theory of meaning does not count as a theory of meaning because it attributes far too little sensitivity and understanding to those who are said to understand a language. My response was that knowing a T-sentence within the appropriate large semantic theory, on the basis of a large body of evidence, does qualify one as understanding the left-hand sentence of the T-sentence. That was an appeal to holism, more or less.

Holism is implicit in the structure of Tarskian theories, and it was explicit from Davidson's first presentations of his theory of meaning:

> If sentences depend for their meaning on their structure, and we understand the meaning of each item in the structure only as an abstraction from the totality of sentences in which it features, then we can give the meaning of any sentence (or word) only by giving the meaning of every sentence (or word) in the language. (Davidson 1967, p. 22)

Of course, this degree of holism is too much: surely I don't have to be able to understand *everything* a speaker might say, given his current speech dispositions (that is, which Tarskian theories will work as interpretive theories of meaning for him) in order to understand *anything* he might say. I return to this issue below, after showing why holism solves the three objections with which I began the section. For the moment, assume that holism has been moderated into plausibility.

Consider the first issue: interpreting beaches. Interpretation consists, according to Davidson, in generating a Tarskian theory for a speaker that assigns its truth-conditions to each possible utterance of that speaker. How might I go about doing this for the beach? If I wanted to interpret each of the footsteps as the utterance, "I was stepped on here," given the nature of semantic theories of truth, I would have to derive, as a theorem, the T-sentence "footprint" means (in the beach's idiolect), as uttered by beach B at time t and

place p that B was stepped on at p at t, the moment at which the utterance began. But from what axioms could I possibly derive this sentence? Only if I forewent the holistic nature of Tarskian theories would I try to have the theorem without its substructure. If we require the T-sentences to be derived from axioms, then we won't try to interpret beaches, or anything else that can't be interpreted on the basis of the axioms of a Tarskian theory.

Recall the enemy radio operator from last section. The radio operator's language, Code, is much like the beach's would-be language, in that each utterance is absolutely atomic, not composed of parts that can be used to relate utterances to one another. Does my refusal to interpret the beach not undermine my use of the radio operator as an example?

Yes and no. Recall that the reason we took for granted that we could interpret the radio operator is that we knew in advance that he was human, and we took for granted that we could interpret humans. In the absence of that assumption, it's not plain that the radio operator could be interpreted: his radio signals themselves would not, I suggest, convince us to think that he is a speaker of a language. That is not a defect thrown up as an artifact of the structure of Tarskian theories; it points the way to a deep truth about thought and language.

Davidson appeals to the holistic nature of Tarskian theories to rule out "counterfeit theories" that assign to left-hand sentences non-interpretive material equivalents, such as the theorem that "snow is white" is true iff grass is green. The trouble with such theorems is that it's hard to see how we can have axioms governing "snow" and "white" from which we could both (1) derive grass's being green as the truth-condition for "snow is white" and (2) derive only true T-sentences. Since the axioms are put to use in deriving many theorems, and many of those theorems must be borne out by experience for the axioms to count as confirmed, it's unlikely that axioms with such wild consequences could be maintained. Since holism is necessary to get the right

interpretation, we have a transcendental argument for holism not unlike the transcendental arguments for externalism I discussed in the last section.

Fodor and LePore want to stop the holism right there, by arguing that this fact about how an interpretive theory should look is only a consequence of the fact that it's a theory for a compositional language, rather than the fact that it's an interpretive theory:

> ...this is a good argument for semantic holism only if the appeal to compositionality really is required to rule out T-theories that entail [counterfeit] theories...; and it's possible to doubt that it is. Indeed, on reflection, it's hard to see how it could be. If it's really only because of the structural similarity between "Snow is white" and "That's snow" that the former means that snow is white (and not that grass is green or that 2 + 2 = 4), then it would seem that there is an a priori argument against the possibility of a noncompositional language. The expressions of such a language, according to this argument, *could not* have determinate truth conditions. We doubt that there could be such an argument. (Fodor and LePore 1992, p. 65)

Fodor and LePore offer the example of two children, one of whom speaks atomic sentences of English (but no sentences with sentential connectives), the other of whom utters what Fodor and LePore suggest are one-word "sentences" under the same circumstances. If the first child is willing to infer "That's cold" from "That's snow," then the second one is willing to infer, say, "Mary" from "Sam" (and is inclined to say "Mary" whenever the first would say "That's cold," and "Sam" whenever the first would say "That's snow.")

Since the second child's "language" is non-compositional, if there were no non-compositional languages, then his "language" wouldn't be a language. But *obviously*, Fodor and LePore say, it is. "After all, whether the child means anything by his utterances presumably depends on the intentions with which he utters them. What a priori argument would show that a child couldn't utter "Sam" with the intention of thereby saying that snow is white?" (*ibid.*, p. 66) This contention is obviously question-begging, since if the child can't be attributed anything with the content that snow is white, then he can't be attributed the intention to say that snow is white; and the denial that the child speaks a language at all will rapidly bring in train the denial that the child has intentions.

Nevertheless, it's not at all obvious what prevents the child from having his one-word sentences.

The first thing to be said about this language is that it will have no sensitivity to time and place. The child would never be able to say "That is white," because, to say something meaning that *that* is white is to say something that means something quite different from something meaning that *that* (other thing) is white. Without a word to ostend with, it's not at all plain that the child could manage to ostend anything. I'll temporarily assume that Fodor and LePore somehow manage to solve this problem and that the child could, if there are no other problems, be taken to utter observation reports.

Recall the second objection to the truth-conditional theory: that it fails to account for the intensionality of the semantic. Recall also the third objection: that it fails to give content to non-observation sentences. Those objections apply with a vengeance to Fodor and LePore's language of one-word sentences. The child who speaks the non-recursive component of English can distinguish between "water is wet" and "H_2O is liquid at room temperature," but how can the child with one-word sentences do so? To interpret that child, there would be *nothing but* correlating circumstance to utterance, but a given circumstance can be described in endlessly many ways, many of which would not accurately reflect the state of mind of the speaker prompted by that circumstance to make an utterance with that circumstance as its truth-condition. As Davidson puts it:

> One way of telling that we are attributing a propositional attitude is by noting that the sentences we use to do the attributing may change from true to false if, in the words that pick out the object of the attitude, we substitute for some referring expression another expression that refers to the same thing. The belief that the cat went up that oak tree is not the same belief as the belief that the cat went up the oldest tree in sight. (Davidson 1982, p. 97)

This intensionality of the attitudes is of course shared by the semantic theory specifying contents of utterances. The problem is that for interpretation to succeed, we want our T-sentences to be as sensitive and informative as possible. To assign its truth-condition to a sentence under *just any old description* of that truth-condition is not

adequate for good interpretation. We want to assign truth-conditions under the sorts of descriptions the speaker has in mind. The language of one-word sentences affords us, at best, truth-conditions, but allows us no finer distinctions. We have no basis on which to choose one description of the truth-condition of the utterance rather than another, so we have no way to respect intensionality. But an interpretation that fails to respect intensionality is a bad interpretation.

It might seem at this point that I have a robustly contradictory approach to intensionality. In discussing Sosa's Fregean argument against semantic externalism in 2.1.2, I was, it might seem, dismissive of the intensional aspect of attributions of contradictory beliefs: beliefs might contradict, I contended, without the believer in the contradictory beliefs being irrational. I identified contradictoriness and hence content with something external to the mind, and hence, one would think, with something semantically extensional. If content is extensional, then I've no right to intensionality.

Though it's not immediately plain why, I *do* have the right to intensionality. The truth-conditional theory of meaning, and its attendant approaches to attitude attributions, can fully respect intensionality. At a very superficial level, I understand the problem of intensionality like this. Some utterances are extensionally equivalent: they may be intersubstituted *salva veritate* in extensional contexts. Yet they are intensionally inequivalent: when those utterances appear in attributions of attitudes and other intensional contexts, they may not necessarily be intersubstituted *salva veritate*. For Fregeans and many others, this failure of intersubstitutability is to be accounted for with reference to a difference in meaning between the utterances, and this move drives a distinction between meaning and extension. By identifying meaning with extension, I block that move.

To distinguish between co-extensional utterances, I must appeal to something other than their meanings. But that's easy, for utterances, and the attitudes they express, have *many* features, not just their meanings. There are also utterances'/attitudes' relations

to other utterances/attitudes: that is, there is also a holistic aspect to these states. Davidson remarks:

> Why doesn't the fact that a horse or a duck discriminates many of the things we do strongly suggest that they have the same concepts we do, or at least concepts much like ours? ...there is little reason to take the suggestion literally. Someone could easily teach me to recognize a planet in our solar system without my having a clear idea what a planet is. A horse can distinguish men from other animals, but if it has a concept of what it is distinguishing that concept is nothing like ours. Our concept is complicated and rich: we would deny that someone had the concept of a man who did not know something about what distinguishes a man from a woman, who did not know that fathers are men, that every man has a father and a mother.... (Davidson 2001a, pp. 136-7)

Consider, for instance, my attribution to someone of the belief that Hesperus is Hesperus. I'm unwilling to, on the basis of this attribution and the semantic identity between 'Hesperus' and 'Phosphorus,' attribute to someone the belief that Hesperus is Phosphorus. It's true that this belief has the same content as the first, so I *may*, if I'm not trying to be particularly discerning, attribute the belief that Hesperus is Phosphorus, and I'll be right. But I won't have captured the speaker's state of mind. The belief that Hesperus is Hesperus lacks certain inferential connections with other beliefs that are partly definitive of the belief that Hesperus is Phosphorus; for instance, if someone believes that Hesperus is hot and that Hesperus is Hesperus, she might not be apt to believe that Phosphorus is hot, even though in believing that Hesperus is Hesperus, she does believe something with the same content as my belief that Hesperus is Phosphorus. Subjective intensionality is a consequence of our differential willingness to draw inferences from beliefs that share content, which is a consequence of the opacity of the objects of thought.

The child in Fodor and LePore's thought experiment is prompted by experience to make a noise, say, "Sam." The circumstance that prompts this utterance is snow's being white. But that circumstance is identical to frozen crystalline precipitation's reflecting equally all bands in the visual spectrum. Whether we attribute to his utterance one of these or another as its meaning is indifferent if we ignore mental context, since they're the

same circumstance. But these two beliefs of ours have different relations to our other beliefs; that's what makes them distinct beliefs. If we can't attribute any one of those beliefs in particular, it's hard to see why we should attribute any belief at all.

I want to shift to another problem for a moment before returning to this one. Fodor and LePore claim that the child is willing to draw appropriate inferences, and would no doubt extend their thought experiment to make the child as sensitive as you like. But how could any utterance of the child ever possibly be attributed any content beyond immediate experience? How could Fodor and LePore respond to the third objection to the truth-conditional theory, when applied to their own approach? What could the child ever do to convey the content "2+2=4," as distinct from any other claim that's always true? In contending that Fodor and LePore could not handle this problem, I hope to show how Davidson's could and hence why holism is necessary.

The problem here is how we could attribute to an utterance a content that is general. General utterances are not prompted by the occurrence, in the immediate environment of the speaker, of a circumstance the occurrence of which makes the utterance true. So it's not possible to identify their content with their cause. Fodor and LePore were straightforwardly identifying content with cause, so they have no means to attribute content to general utterances.

That is not a problem that we face with a Davidsonian theory. In a Tarskian theory discovered and tested in the situation of radical interpretation, while we *begin* to assign axioms to words on the basis only of their appearance in observation sentences, we can check our axioms against theorems about general sentences, as well as further observation sentences. If our axiom for some symbol is that something satisfies that symbol iff it is a father, then our axiom is wrong if it mispredicts the role of that symbol in general sentences about fathers. In this way, we can both begin to understand the natives' general utterances, and also confirm and confute our tentative axioms for words appearing in observation sentences.

The child speaking Fodor and LePore's language, though, cannot display the relations between his observation sentences and his more general sentences. We might see something that we suspect to be an inference, but we won't be able to figure out what the inferred general sentence said, only, at best, that it was one of the many general sentences that may be inferred from the observation sentence with which we began. Thus "snow is cold" is beyond reach of the child, as it is general.[41]

Let me return to the second problem. Before we're willing to attribute to someone the belief that, say, snow is white, we would feel the need to attribute to him, among many other beliefs, the beliefs that white is a color and that snow falls from the sky: in general, we would think it necessary to attribute some appropriate set of other beliefs involving these concepts. Some of these beliefs are general in nature, not observation sentences. If the child says something that we should interpret as meaning "snow is white," then we should interpret the child as willing to say that white is a color and that snow falls from the sky. But this we would never do, since the child could never say anything general. So we should never interpret anything the child says as meaning "snow is white." Contrary to Fodor and LePore, I think that we are in possession of an *a priori* argument that there could be no non-compositional language: Observation sentences acquire their identity only in relation to both observed phenomena and general sentences; general sentences acquire their content only by their relation to observation sentences; in a non-compositional language, there would be no appropriate relation to observation sentences; hence there would be no general sentences, and no observation sentences.

Davidson's view is not idiosyncratic. Sellars, for instance, agrees:

> ...one couldn't have observational knowledge of *any* fact unless one knew many *other* things as well. ....the point is specifically that observational knowledge of any particular fact, e.g. that this is green, presupposes that one knows general facts of the form *X is a reliable symptom of Y*.

[41] Likewise, of course, the equally general 'Snow is white,' but I give that example to Fodor and LePore.

> The essential point is that in characterizing an episode or a state as that of *knowing*, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says." (Sellars 1956/1997, pp. 75-6)

Recall that earlier, in 2.1.1, I distinguished between cases where attitude attributions on the basis of mere causal correlation between an utterance and some prompting circumstance could seem warranted, and those in which such attributions were obviously unwarranted. It might seem that, even in the absence of some supporting attributions, it's reasonable to attribute to someone a belief that can be expressed as a simple observation report. Such beliefs are prompted by immediate experience. But if Sellars is right, these cases also require attributions of additional, supporting content. The point, though, of noting Sellars's view is just that Davidson's view is actually widespread, not just one more unusual feature of an unusual system. Many Kantians who try to understand attributing content as attributing rationality will share some form of holism.

By showing how Fodor and LePore could not handle the second and third objections, I hope to have shown how Davidson can. By insisting on using Tarskian theories, with their essential holism, in interpretation, a Davidsonian interpreter could respect the intensionality of utterances and attitudes, and could come to assign general content to various beliefs and sentences.

However, the holism with which Davidson began was overkill. Fodor and LePore explain:

> ...*if* you assume that properties like *having a meaning in L* and *having the same meaning as some expression in L* and the like are holistic, then a certain standard picture of how communication and language learning work would seem to be in jeopardy. The picture is that the linguistic and theoretical overlaps of speaker and hearer can overlap *partially* to any degree you like: you can believe some of what I believe without believing all of it; you can understand part of my language without having learned the rest of it, and so forth. This would seem to be essential to reconciling the idea that languages have an interpersonal, social existence with the patent truth that no two speakers of the same language ever speaker exactly the same dialect of that language. (Fodor and LePore 1992, p. 10)

If it were impossible for utterances in slightly deviant dialects to share content, then no two speakers could ever agree or disagree about anything, which would be bad for our understanding of scientific and moral progress — to say nothing of being bad for scientific and moral progress.

Luckily, no holism so extreme is warranted by the nature of a Tarskian theory, as Davidson notes: "We could not recognize as capable of thought a mind that did not conceive of a supply of familiar objects and properties. Just which objects and properties *is not fixed*, though no doubt there are some we could not do without." (Davidson 1995, p. 13, emphasis added) If the objects and properties are not fixed, they are not fixed to exactly the same objects and properties of which I conceive. So Davidson rejects the extreme holism that he confusedly announced in the earlier paper.

Fodor and LePore are not inclined to let holism off so easily. They discuss[42] the possible response to their argument that runs like this. Holism commits us to an attitude/utterance's identity being determined by context. Unfortunately, no two such contexts — minds, idiolects — are identical. So no two attitudes/utterances are identical. Nevertheless, they could well be *similar* — in triggering circumstance and inferential relations — and that's all we need.

But this is plainly inadequate. There's no plain way to apply the concept of similarity of triggering circumstance or inferential relation when there's in principle no way to apply the concept of identity of trigger or inference, even as an idealization. But if identity of belief required identity of inferential relations, then two minds would have to have exactly the same beliefs in order to share any beliefs at all. Trying to push similarity of belief off onto similarity of something else requires that the something else be identified independently of beliefs, but of course inferences are identified with reference to their premises and conclusions and the rules of inference employed in them, and the premises and conclusions are beliefs.

---

[42] Fodor and LePore 1992, pp. 17-22

There is no reason, though, why Davidson could not have identity of utterance or belief. The holism that requires identical minds for identical beliefs is too extreme. We can have identity of content in merely similar mental contexts. To solve the problem, we need to clarify what it is for attitudes to be identical, and we need to focus on the precise balance of externalistic and holistic aspects of interpretation.

Two attitudes have the same content just in case they share truth-condition, but sharing content is inadequate for two attitudes to be identical. They also have to (correctly) seem to the person bearing the attitudes to warrant many of the same inferences and actions. In using a Tarskian theory to interpret, we try to give axioms that not only capture the truth-conditions of our interlocutor's utterances, but also identify an utterance within a web of other utterances, potential and actual, that help define its role in the rational life of the interlocutor. Two rational lives need not be entirely identical in order to be similar enough to share content. I think that this might best be illustrated by an extended example. The point of this example is that it proceeds according to the method of radical interpretation, and the holism is taken for granted; nevertheless, we won't find ourselves being tempted by the implicit holism to deny identity of content between two agents. That is, I'm not just going to show that we are not extreme holists in our interpretive practice, which is obvious. I'm going to show that we are not extreme holists when our interpretive practice is holistic in the way that Davidson's Tarskian radical interpretations must be.

Assume that we have hypothesized the axiom for our theory of speaker s that '🍽', as uttered by s at time t, refers to the object of s's ostension. This axiom seems to help us generate many well-confirmed theorems. While ostending a patch of snow, s utters "🛠🍽." Also, when snow begins to fall, s utters "🛠🌑." We might imagine that the former means "That's snow," and the second means "Snow falls." We could derive these T-sentences as theorems were we to posit the axiom that for any x, x satisfies '🛠,' as spoken by s, iff x is snow. (And an appropriate axiom for '🌑.')

Assume that we continue in the interpretive endeavor, and find that we can get good confirmation on the posited axiom that for any x, x satisfies '🏝,' as spoken by s, iff x is cold. When we inquire whether "🏝🛠," we get what we have theorized to be an affirmation. That encourages us in our axioms, since s seems to have agreed that snow is cold, which our axioms predict she would say. If s were more inclined to agree to the sentence the more cold snow she had seen, then we would feel even more encouraged.

But now let's assume that s also tells us that "🌛🌕," while ostending the moon. We hypothesize the theorem that "🌛🌕," as uttered by s, is true iff the object of ostension is the moon, and we formulate the axiom that '🌛,' as uttered by s, refers to the moon. (Devices to display the identity predicate, and predicative structure in general, seem either to be absent or to be escaping us; perhaps they're implicit?)

But we then ask whether "🌛🏝," and we get bewilderment. The question seems to confuse s, whose people have never been the moon and don't know that it's cold. Should this lead us to back away from our posits about the reference of '🌛' and satisfaction conditions of '🏝?' This unexpected confusion at a well-known truth *does* provide evidence against our axioms. But it's very weak evidence. We can explain the confusion either with reference to theory error, or with reference to s's ignorance. We can't appeal to s's ignorance every time a prediction fails, but we can do so once in a while, especially where ignorance seems a plausible account for an otherwise outlying utterance.

Assume that we find ourselves in this position. It's incredible to suggest, as Fodor and LePore do, that the holistic nature of the semantic theory and attitudinal attributions would lead us to reject our entire theory and abandon any hope of identifying s's meanings and attitudes with mine. Surely there is enough common ground to allow for identity by any reasonable standard thereof. S's utterances that I take to mean that snow is white and so forth play sufficiently similar roles in her mental life and mine for the correlation between them to function interpretively as an identification of s's meanings and beliefs. Fodor and LePore point out, correctly, that:

> ...we need to know *how much* the differences between the red-inferences I endorse and the ones that Shakespeare did count as differences in our concept of red. The extent to which this sort of question lacks a principled answer is the extent to which we have no notion of similarity of content that is compatible with a holistic account of belief attribution. And it lacks a principled answer entirely... (Fodor and LePore 1992, p. 21)

I agree that there is no principled answer to the question of how much divergence there can be between two minds/languages while there can still be shared content. There can be some; that's obvious. As I plan to show in the next chapter, there can't be a lot. But the lack of a principled answer, while regrettable, does not vitiate holism. Making the appropriate judgments is easier in practice than in theory: we make the judgments all the time without having any very good theory about how to do it, but, as I argued above, languages have to be compositional and hence have to be holistic.

I want to address a final problem, an apparent tension between holism and externalism. De Rosa argues:

> According to [Davidson's holism], the pattern of sentences and beliefs embedded in a language determines the meaning and content of each sentence and belief of that language. ...Davidson's externalism... says that in the case of occasion sentences, their meaning and the content of the beliefs they express are determined by some sort of causal relation between tokens of these sentences/beliefs and extra-representational events in the world.... Externalism would readmit two theses that holism ruled out, namely:
>
> (a) atomism, that is, the view according to which a belief or a sentence in a language can have the meaning or the content it has independently of any pattern of beliefs and sentences in which it is embedded; and
>
> (b) (a radically non-epistemic) realism, that is, the view according to which there is a world totally independent of our beliefs such that it could reveal all our beliefs about the world to be false. (De Rosa 1999, p. 203)

On the face of it, the challenge is odd. How does externalism "readmit" theses ruled out by holism? Presumably, the conjunction of externalism and holism rules out everything that was ruled out by either conjunct. I guess that De Rosa has in mind that externalism's "readmitting" theses (a) and (b) consists, not in not implying their negation — just being consistent with them — but in implying them.

Hopefully, I've made clear how (a) is to be dealt with. An utterance's content is its worldly truth-condition. But to get refined T-sentences that really help us get into the mind of someone we want to interpret, we want to give a sentence its truth-condition under a description suited to the way our interpretee thinks of it. That requires that we attend to mental context. Externalism says that content is external, not that factors internal to the speaker have no effect on what external entity is to be the content, or how we should describe that entity when attributing it as content. So externalism does not imply (a).

(b) is a typically overblown statement of realism, one that pretends that the point of realism is to imply scepticism. As I argued in the section on truth, truth-conditions do exist; they are not mere theoretical constructs. So I accept the part of (b) that states realism ("there is a world totally independent of our beliefs"), even while I reject the pretended epistemic point of realism (the independent world "could reveal all our beliefs about the world to be false"). I think that De Rosa is entirely right to point out that externalism implies realism, but whereas Davidson tried to wiggle out of the consequence, I accept it.

De Rosa's challenge is that externalism implies (a) and (b), while holism implies their negations. Externalism does not imply (a), so there is no problem there. Externalism does imply (b). But why would holism imply (b)'s negation? De Rosa's argument here[43] relies on the assumption that holism requires identity of mind/language for identity of any content anywhere. Since that condition is never met, there is never identity of content between any two speakers. There can be no content over which two speakers can disagree. The role of the external world in resolving disagreement, then, which is central to realism, drops out.

This argument is uncompelling. Realism is the claim that the truth of one's utterances/beliefs is determined by their content and how things stand in the world.

[43] De Rosa 1999, p. 203-4.

Realism, on its own, doesn't require disagreement to be possible. Again, De Rosa is confusing a certain epistemic implication (or pretended implication) of realism with realism itself. If we couldn't disagree, then realism would still be true. Whether my beliefs are true would still be determined by meaning and world. It's just that we would never have any disagreements to refer to the world for adjudication.

The argument is also uncompelling because it takes for granted that the holism to be adopted is of the most extreme kind available, one that makes identity of entire mind/language a prerequisite for any shared content. As I've tried to make clear, that sort of holism is not implied by the structure of a Tarskian truth theory, so it is not implied by Davidson's theory of meaning. A more modest holism is all that's called for.

2.4 Semantic and Attitudinal Indeterminacy

Theses like the inscrutability of reference, the indeterminacy of translation or interpretation, and the relativity of ontology are ugly consequences, if consequences they are, of abiding by the publicity constraint in theories of meaning. To go personal for a moment, these sorts of theses freak me out. I rather wish that Quine had never thought of them, and I wish that Davidson's discussions thereof weren't such sterling examples of philosophical obscurity. But for good or ill, a modest indeterminacy of interpretation is a consequence of going empirical in the theory of meaning. Since I don't know of any reasonable alternative to going empirical in the ways described in previous sections, I don't know of any reasonable alternative to accepting a modest indeterminacy. In this section, I discuss semantic and attitudinal indeterminacy, trying to show its scope and limits. The indeterminacy has two sources: the structure of Tarskian theories, and the nature of radical interpretation. Davidson puts out the first sort of indeterminacy here, along with some mistakes:

> We don't need the concept of reference; neither do we need reference itself, whatever that may be. For if there is one way of assigning entities to expressions (a way of characterizing 'satisfaction') that yields acceptable results with respect to the truth conditions of sentences, there will be endless other ways that do as

> well. There is no reason, then, to call any one of these semantical relations 'reference' or 'satisfaction.' (Davidson 1977, p. 224)

I want to discuss everything after the first sentence first, and then return to the first sentence, which is false. This passage points out a truism about empirical theories with the general form that Tarskian theories share. Such theories will consist in a set of axioms that are not themselves testable. The axioms will yield a set of theorems that are more or less testable against observation. But for any such set of theorems, if there is a set of axioms that imply them, there are many more such sets of axioms that would imply them. Thus, the evidence underdetermines the theory: many theories are consistent with all of the available evidence. In the case of Tarskian theories, there are many sets of axioms stating the reference and satisfaction conditions of individual words of the language that will yield the right theorems.

It is often said that this is a less-than-profound and totally familiar instance of the underdetermination of theory by evidence, as, for example, by Searle:

> It is only by *assuming* the nonexistence of intentionalistic meanings that the argument for indeterminacy succeeds at all. Once that assumption is abandoned, that is, once we stop begging the question against mentalism, it seems to me that [the] objection [that what Quine has demonstrated is mere underdetermination] is completely valid. Where meanings psychologically construed are concerned, there is the familiar underdetermination of hypothesis by evidence, and that underdetermination is in addition to the underdetermination at the level of physical particles or brute physical behavior. So what? These are familiar points about any psychological theory. There is nothing special about meaning and nothing to show that where meaning is concerned there is no fact of the matter. (Searle 1987, p. 232-3)

What this objection fails to grasp is that, with language, all is public. There are no semantic facts that can't be grasped on the basis of publicly available evidence. So, since the publicly available evidence leaves the semantic facts indeterminate, the semantic facts are indeterminate. Nothing beyond the public use of language can determine what theories of meaning are right about the user of the language, and public use of language doesn't determine that to the point of uniqueness. Thus there is no unique theory that is

right about the speaker. Searle claims that he accepts the publicity constraint, but he doesn't see what it implies:

> ...let us grant that, for "public" languages... there is at least one clear sense in which semantic features are, indeed, public features. I take it all that means is that different people can understand the same expressions in the same way.... Furthermore, let us grant, at least for the sake of argument, that the public features are subject to underdetermination in at least this sense: I could give different but inconsistent interpretations of someone's words, all of which would be consistent with all of the actual and possible evidence I had about which sentences he held true. Now what follows? ...on Davidson's view the indeterminacy follows only if we assume from the start that different semantic facts must necessarily produce different "publicly observable" consequences. Only given this assumption can we derive the conclusion that a speaker's meaning and reference are indeterminate and inscrutable. (*ibid*, p. 244)

By Searle's notion of what it is for semantic features to be public, we must all be able to understand the same expression in the same way. If the semantic facts did not produce different publicly available consequences, there would be no public basis by which we could coordinate our utterances and all mean the same thing by the same utterance. Whatever mentalistic phenomenon I attach as meaning to my utterance would be beyond being grasped by you, so you could never figure out what I meant. You might get lucky, and happen quite by chance to attach, to utterances of yours similar to utterances of mine, private mental meanings similar to the ones I attach to utterances of mine. But the odds of that happening are vanishingly small; small enough that, if that were the condition on understanding the same expressions in the same ways, we would never do it. While Searle is right to point out that indeterminacy is a consequence of taking a third-person perspective in theory of meaning, whatever we have access to only from the first-person perspective is not public, and hence not meaning. Indeterminacy is thus a consequence of taking the only perspective on meaning that we can.

Searle makes other, more local, errors. He contends that one consequence of the indeterminacy would be that we can give "inconsistent" interpretations. Elsewhere, he says that, "The thesis of the indeterminacy of translation is that, where questions of

translation and, therefore, of meaning are concerned, there is no such thing as getting it right or wrong. This is not because of an epistemic gulf between evidence and conclusion, but because there is no fact of the matter to be right or wrong about." (*ibid*, pp. 230-1) But it's true neither that the inscrutability of reference allows us to offer inconsistent interpretations, nor that the inscrutability of reference implies that we can't give a bad interpretation.

The claim that inconsistent interpretations can be given is repeated by other commentators; Ludwig and Lepore, for instance, contend that, "...the interpreter must regard the different theories he can confirm as *strictly* incompatible with one another. According to the two theories, sentences of the object language will mean different things. In other words, it is incoherent for the interpreter to regard the different theories which he could confirm as both true." (Ludwig and Lepore 2005, p. 239) These accusations miss the point of Davidsonian indeterminacy. Davidson says:

> ...we can suit the evidence by various ways of matching words and objects. The best way of announcing the way we have chosen is by naming the language; but then we must characterize the language as one for which reference, satisfaction, and truth have been assigned specific roles. An empirical question remains, to be sure: is this language one that the evidence permits us to attribute to this speaker? (Davidson 1979b, p. 240)

We may think of a language as defined by a Tarskian theory. For any given set of theorems of such theories, there are many sets of axioms from which they could be derived. Each such set of axioms, along with the common set of theorems, is a distinct language. They are, nevertheless, empirically equivalent. Since there are no semantic facts not knowable on the basis of publicly available evidence, if two theories are empirically equivalent, either of them define a language that a speaker could be said to speak if she could be said to speak the other. But that doesn't imply that *just any* language could be attributed to *just any* speaker. Only the ones defined by Tarskian theories with the right theorems can be attributed to her.

Furthermore, there can be no inconsistency in the attributions. Consider how the indeterminacy would work. I could, let's say, give either of the following interpretations of some speaker's utterances, depending on whether I treat her as speaking L_1 or L_2:

'Apollo' refers-in-L_1 to Apollo

'Apollo' refers-in-L_2 to Lucky

There's plainly no inconsistency. That a word refers to one thing in one language is not inconsistent with its referring to something else in another language. Likewise, there is no inconsistency at the level of theorems:

'Apollo is sleeping' is true-in-L_1 ↔ Apollo is sleeping

'Apollo is sleeping' is true-in-L_2 ↔ Lucky is sleeping

Having one set of truth-conditions in one language is consistent with having a different set of truth-conditions in another language.

At this point, though, we should begin to rein in the indeterminacy. For radical interpretation to succeed, our interpretive theorems must abide by the version of the principle of charity known as Grace, according to which the right-hand side of the T-sentence must give the cause of the utterance of the object-language sentence under discussion. This fact rules out many possible permutations of reference and satisfaction that would yield theorems that gave mere metalanguage material equivalents of object-language sentences. It would be very hard to come up with seriously divergent axioms of reference and satisfaction that yielded equally satisfactory theorems for observation sentences of the object language.

Meditation on this point allows us to undo the standard 'gavagai' example of indeterminacy. Let's assume that I have generated the following theorem for one of my informants:

'Gavagai!' is true-in-L_1 for a speaker s at time t ↔ a rabbit is present near s at t

Quine wants to claim that the following T-sentence would do equally well:

'Gavagai!' is true-in-L_2 for a speaker s at time t ↔ an undetached rabbit part is present near s at t

Since the second theorem is empirically equivalent to the first, we could, Quine says, equally well say that the speaker is speaking L_2 as L_1. And, since a rabbit being in one's presence is nomically equivalent to an undetached rabbit part being in one's presence, it might seem hard to distinguish between the two and end up identifying the correct cause.

Speakers are sensitive to events. But, due to the opacity of events, events might have many features to which the speaker is not sensitive; that is, events might have features with no, or no salient, nomic correlation to the speaker's utterances. In identifying the cause of an utterance, we want not only to identify the relevant event, but also to identify it under a revealing description. The first desideratum here is that we get a description that we can relate to the utterance as a matter of law. A rabbit's being present is also a mammal's being present, but a mammal's being present does not have a lawlike relation to utterances of 'Gavagai!' That's why Quine's examples are all of other events *nomically* equivalent to the rabbit's being present.

The second and more important desideratum is that we try to narrow in on descriptions that describe the event *as the speaker understands her sensitivity to it*. That is, we want our T-sentences to, as much as possible, state laws that the speaker would affirm. This desideratum is only the desideratum discussed in the section on holism: that we try to make our T-sentences reflect the states of mind of those we interpret as closely as possible, that we respect the intensionality of their utterances.

The indeterminacy that remains will be quite limited. When it comes to observation sentences, the indeterminacy would probably be vanishingly small unless there are serious divergences between the resources of the object- and meta-languages. For the standard example, 'Gavagai!,' for instance, since I can tell perfectly well when someone is talking about rabbits and when she is talking about undetached rabbit parts, I could across time narrow in on the correct description of the event of the rabbit's being

present that describes the event in the most revealing way possible. Only if we genuinely couldn't tell which description(s) of the event the speaker felt himself to be sensitive to the event under, would we have indeterminacy at the level of observation sentences. We genuinely couldn't tell what the speaker had in mind only if the speaker could never say anything to make the distinction between rabbits and their undetached parts. But if he makes the distinction at all, he can exhibit that fact.

Davidson argues that the indeterminacy, as I have described it, is no more threatening to the objectivity of meaning assignments than the difference between Fahrenheit and Celsius scales is to the objectivity of temperature assignments. Lepore and Ludwig reply[44] that this defense would be devastating to the project of radical interpretation. The use of numbers to keep track of temperatures allows for multiple scales — assignments of numbers to temperatures — only because numbers are "richer" than temperatures: "The possibility of [keeping track of an empirical pattern among temperatures] in different ways shows that the pattern of relations among numbers is richer than that among the states that the number are used to keep track of." (Lepore and Ludwig 2005, p. 245) In general, if some phenomenon x can be represented, in multiple scales, by reference to distinct phenomenon y, then y is "richer" than x. However, when the radical interpreter interprets her own utterances, the language to be interpreted is not richer than the language in which interpretation occurs. Thus there could be only one scale.

It seems to me that, if we assume that Lepore and Ludwig mean something clear by "richer than" — which is not obvious to me — then this is just correct and unproblematic. The indeterminacy of meaning is the fact that any language can be interpreted in the terms of another language in more than one way, not that any language can be interpreted in the terms of any language, even itself, in more than one way. They disagree, and contend that their point about richness shows that, as a background

[44] Lepore and Ludwig 2005, pp. 245-7.

assumption for interpretation, an interpreter must accept that "The relations among the object language sentences are not as structurally rich as those among the sentences of the interpreter's language." (*ibid*, p. 246) But this assumption seems to come out of nowhere. If we insist on the possibility of indeterminacy as a prerequisite for interpretation, then we might have to accept that the object language is poorer than the meta-language (though in fact we wouldn't; see below). But why insist on indeterminacy? Indeterminacy is a *side-effect* of going empirical, not something we went empirical to achieve.

Even in object languages of equal or greater richness, though, there would still be a modest indeterminacy. For more theoretical sentences, there will be more interpretations available. The closer the ties between the sentence and prompting experience, the narrower will be the indeterminacy, because it will be correspondingly harder to come up with multiple descriptions of events all of which could, with equal plausibility, be said to be descriptions of events that the speaker would accept as giving the description under which she is sensitive to the event. Only in districts of theory with very little tie to the empirical is there apt to be really wide indeterminacy. There is, thus, just enough indeterminacy of meaning to be consistent with the nature of empirical theories, but not enough to seriously undermine our sense of ourselves as speakers whose utterances have a certain definition to them.

I mentioned the second desideratum, flowing from holism and a desire to respect intensionality, of assigning the content event to observation sentences under a description under which the speaker thinks of herself as sensitive to the event. Identifying such descriptions would, of course, be enormously difficult. It leads to the second source of indeterminacy, the nature of radical interpretation. I can deal more quickly with this source of indeterminacy, since major objections have been covered already.

In radical interpretation, I seek to interpret a subject. That consists in identifying the meanings of her utterances and the contents of her beliefs. To recur to the standard analogy, radical interpretation tries to solve one equation with two variables. But of

course that can't be done, at least not to the point of uniqueness. We're left with a line through a coordinate grid, in which meanings are represented along one axis and beliefs along another. Infinitely many assignments of meanings and beliefs can be rendered consistent with the speaker's utterances. The idea is that, if an attribution of some belief seems odd, we can change it if we make compensating adjustments to assignments of meanings; if a certain meaning seems hard to assign, we can assign a different one by compensatory shifts to belief attributions. Davidson says:

> Underlying the indeterminacy of interpretation is a commonplace about interpretation. Suppose someone says, 'That's a shooting star.' Should I take him to mean it really is a star, but that he believes some stars are very small and cold; or should I think he means it is not a star but a meteorite, and believes stars are always very large and hot? Additional evidence may resolve this case, but there will always be cases where all possible evidence leaves open a choice between attributing to a speaker a standard meaning and an idiosyncratic pattern of belief, or a deviant meaning and a sober opinion. (Davidson 1973, p. 257)

Grace sharply restricts the available attributions, of course, so what we have is not so much a line as a segment. But surely there is real, if limited, indeterminacy here.

But we can ask just how much Grace can restrict, given the notion that only in triangulation and radical interpretation does content come into being. In 2.2, I answered a number of questions about the assignment of content in triangulation and radical interpretation. One question that went unanswered was, Couldn't the observer assign the wrong point along the line? Instead of assigning the Ψ events as causes of my Φ responses, what if she were to assign the causes of the Ψ events? The point of the objection is that it seems as though an observer could go wrong, and if an observer could go wrong, then there is an independent standard for determining content, one which lends determinacy to content.

The objection misplaces the standard. Of course there is a standard, in the sense that an interpreter can make mistakes. But if the "mistake" appears in a well-confirmed empirical theory for a speaker, consistent with all the facts (that there will ever be), then it isn't a mistake. Absent the interpreter, no point along the line counts as right or wrong,

so it makes no sense to suggest that an interpreter could go wrong *if her theory is empirically adequate*. But, since there can be multiple empirically adequate theories, then, if some of them assign different points along the causal chain as causes, then all of those points equally well count as causes and hence as contents.

The main significance of interpretation-based indeterminacy, from my point of view, is that it reminds us of a basic Davidsonian truth: "Having a language and knowing a good deal about the world are only partially separable achievements..." (*ibid*) In the situation of radical interpretation, I solve simultaneously for meaning and belief. Meaning and belief, then, can be seen as functions of one another. Much of what we want to say about meaning, then, will apply across this functional relationship to belief. For instance, we warp semantic assignments to allow for better attitudinal attributions; but the content of the utterance getting an assignment is the same as the content of the attitude we attribute on the basis of the utterance. If the content of the utterance, then, is its cause, then the content of the belief is the same external event. If we consider a number of previous semantic assignments before making an attitude attribution, the holism of language has played across into a holism about belief.

2.5 SELF-KNOWLEDGE

Attitudes are objects of knowledge; it's possible for me, and for you, to know that I believe that such-and-such. At a minimum, this fact gives rise to a philosophical *question*; that question will lead to problems or paradoxes depending on one's approach to belief and content.

Consider the sort of classical Cartesian view of the attitudes and their content that people like to accuse other people of still accepting. According to this internalist approach, what it is for a belief to have a certain content is, if not to have a certain phenomenology, at least determined by its phenomenology. So there is no question of my knowledge of my own beliefs: I can feel the way the belief feels, and there's nothing else I need to know in order to know the content of the belief. This route permits the other

minds problem, since the phenomenology of other minds is something to which outside observers have no access and about which sceptical doubts can be raised.

Externalism, though, when combined with the notion of first-person authority, gives rise not just to a problem, but to a paradox. For Davidson, what an interpreter can discover about my inner states is what there is to know about them. As a consequence of the fact that content only emerges in the context of triangulation, what an interpreter can discover is *the standard* for content. So other people's knowledge of my attitudes is certainly accounted for. But what about my knowledge of my own states? Since I can't triangulate for myself, how can I grasp the contents of my own thoughts? How can I know my own mind?

So far, this is only a problem, on a par with the problem of other minds but reversed. We can be confident that the problem will find solution, since we're absolutely convinced that we know what we're thinking. But that conviction leads to paradox. For we think that, not only do we know what we're thinking, but also that we have special authority when it comes to our own attitudes: our beliefs about our attitudes are, if not infallible, very close. They're much closer to infallible than any outside observer's. So, despite the fact that the interpreter's point of view provides the standard for content, my own point of view is more reliable than the interpreter's. That's paradoxical. How could the standard be less reliable than anything else?

Paul Boghossian offers a *reductio* of externalism that looks a little bit like an externalism anti-sceptical argument:

> ...let us suppose that Oscar... is a compatibilist [about externalism and authoritative self-knowledge]. I claim that Oscar is in a position to argue, purely a priori, as follows:
>
> (1) If I have the concept <u>water</u>, then water exists.
>
> (2) I have the concept <u>water</u>.
>
> Therefore,

(3) Water exists.

> Since the conclusion is clearly not knowable a priori, one of the premisses in Oscar's evidently valid reasoning had better be either false or not knowable a priori. (Boghossian 1998, p. 275)

(1) is intended to express one consequence of externalism, one similar to the premises of the anti-sceptical argument of 3.2.3 that I will label synthetic *a priori*. Our confidence in (2) is intended to be one consequence of the authoritativeness of self-knowledge. For Boghossian, each premise is known *a priori*, and that's the problem. If two premises are each known *a priori*, and they imply some conclusion, then that conclusion can be known *a priori*. But, though (3) is implied by the premises, (3) isn't knowable *a priori*, so the premises aren't each knowable *a priori*.

The natural approach to take is to wonder whether the premises are, in fact, knowable *a priori*. To suggest that self-knowledge is *a priori* must be judged eccentric. That I know that I have a certain belief only on the basis of *inner* experience hardly implies that I know that I have that belief on the basis of *no* experience. Boghossian retreats from, or rephrases, his point: "...the a priori knowability of premise (2) just *is* the view that I have called the doctrine of privileged self-knowledge..." (*ibid*) But *a prioricity* and privilege are not the same. If the complaint is that (3) can be known with *certainty*, on the basis of an anti-sceptical argument based on the nature of content, then the complaint is just the insistence that the sceptic not be defeated by an externalist argument. I assume, though, that Boghossian can find some suitable description of the premises that neither begs the question nor is obviously confused. I should not be able to tell that water exists by engaging in philosophical reflection and personal introspection.

The argument is directed at someone who can tell, by introspection, that she has the concept of water. That's because Boghossian directs the argument at Putnam style externalism, based on the Twin Earth thought experiment and the notion of a natural kind. Putnam's (and Kripke's) externalism involves the sort of term-by-term determination of content that Davidson rejects. For Davidson, concepts are theoretical

constructs, abstracted from beliefs. Identifying myself as a holder of the concept of water requires that I identify myself as a believer about water, and then abstract the concept from the beliefs of which it is a constituent. So the immediate self-knowledge is of my beliefs, not the concepts that I abstract from them. But what sorts of beliefs could I have from which I could abstract the concept water but that *don't* already commit me to believing that water exists? Furthermore, *those* beliefs will all be *a posteriori*. So Boghossian's *reductio* reduces to someone deducing, from his belief that water is wet and (the result of the anti-sceptical argument) that most of his beliefs are true, that water is wet. This does not represent any further cognitive achievement beyond the one already performed: acquiring the belief that water is wet. No new belief comes into the system through engaging in the *reductio*'s reasoning. So no new *a priori* belief comes into the system. Boghossian's *reductio* fails.

But that doesn't solve the main problem, which is the paradox of externalistic self-knowledge. Why not just purchase the anti-sceptical result of the next chapter with a loss of self-knowledge? In a way, I wouldn't mind that result. I'm a lot less interesting than the rest of the world, and however self-obsessed someone is, surely very few of his beliefs are about himself. So trading self-knowledge for secure knowledge of the world would be a good trade.

Nevertheless, it's impossible. It's easier to go wrong about the world than it is about one's own attitudes. In interpreting, I would have to go wrong if I were to systematically interpret someone as believing that ϕ but also believing that she does not believe that ϕ. Such an interpretation would make a hash of the speaker's self-attributions and/or her first-order beliefs; anyone so badly confused about her own attitudes doesn't count as having attitudes. This fact is a consequence of the holism of the attitudes. But, as Barry Smith points out, that fact is not enough to solve the paradox:

> Has Davidson explained how thinkers know what they are thinking without interpreting themselves? Given that there is something the speaker means when he holds a sentence true, there is something he believes. But in order to *know*

> what he believes, surely he has to *know* what he means. *We* know this by constructing an interpretative truth-theory for his language that assigns truth-conditions to the sentences he holds true. But how does *he* know what he means? As an account of self-knowledge, it tells us *that* the speaker knows what he believes, but not *how* he does. There is the interpretatively guaranteed fact of meaning something, or of our *words* meaning something. But what kind of knowledge does this give us of what we mean and think? (Smith 1998, pp. 416-7)

Knowing that my higher-order self-attributions are massively correct doesn't tell me why they are.

Here is a model of self-knowledge. Self-knowledge is formally very much like interpretations of others. I self-attribute attitudes toward certain truth-conditions, using sentences, and I understand those sentences, thereby giving substance to the self-attribution. What we may think of as belief in certain sentences is the expression of self-knowledge, but "I believe that the sky is blue" doesn't count as self-knowledge unless I understand the embedded sentence. We may express that knowledge in the familiar Tarskian form, "'The sky is blue,' as uttered by me now, is true iff the sky is blue."

When I say that "I believe that ϕ," the word 'that' might refer to the sentence, 'ϕ.' Or it might refer to 'ϕ's truth-condition, the "fact" that ϕ. In either case, the self-attribution only counts as self-knowledge of I can connect 'ϕ' with the "fact" that ϕ. When I say that "'ϕ,' as uttered by me now, is true iff ϕ," I connect the sentence directly to its truth-condition. Self-knowledge is to be knowledge of content, but content is how the world is if my beliefs are true. Self-knowledge, then, is knowing how the world is if my beliefs are true. Such knowledge can be grasped in the form of homophonic T-sentences for my own attitudes.

Having self-knowledge homophonically might make things seem too easy:

> ...self-knowledge is both fallible and incomplete. In both the domain of the mental and that of the physical, events may occur of which one remains ignorant; and, in both domains, even when one becomes aware of an event's existence, one may yet misconstrue its character, believing it to have a property it does not in fact possess. How is this to be explained? I know of no convincing alternative to the following style of explanation: the difference between getting it right and failing to do so (either through ignorance or through error) is the difference between

> being in an epistemically favorable position with respect to evidence — and not. To put this point another way, it is only if we understand self-knowledge to be a cognitive achievement that we have any prospect of explaining its shortcomings. (Boghossian 1989, p. 167)

The model that I suggested above seems to simplify matters too much. Could "Know thyself" have possibly been the substantive command Socrates treated it as, if knowing thyself was so easy?

While self-knowledge is hard, there are two aspects of attitudes that could account for the difficulty. Perhaps knowing the contents of my attitudes is hard. Not so, I say, since having the attitude *is* knowing its content; i.e., its truth-condition, plus believing, desiring, or whatever that content.

Perhaps knowing *the attitude* is the problem. Grasping which sentences you have attitudes toward, and what your attitudes toward them are, is not always easy. I think that I believe that so-and-so is a good person, but in truth, I don't: I believe that so-and-so is a bad person. I thought that I desired that such-and-such occur, but in truth, I don't: I desired that some contrary, this-or-that, occur. The cognitive achievement of self-knowledge lies in grasping what attitudes you have, not what they're attitudes toward. Since externalism is a thesis about content, not about the structure or nature of the attitudes themselves (except insofar as they're structured around external conditions), externalism makes this aspect of self-knowledge neither easier nor harder than it should be.

What about the paradox? It's strange that, while what an interpreter can discover is the standard of content, an interpreter is much less liable to make errors about his own content than an interpreter. There are two things that need saying.

First, "an interpreter" is a bit of an abstraction from all of the other people who interpret one. The interpreters as a group help determine the content of one's thoughts, by triangulating with one. They can't, in general, go wrong while still interpreting. But any individual one of them can go wrong (and, occasionally, they could even all go wrong,

though their interpretation will begin to lack coherence at that point). The authoritativeness of self-knowledge consists in the fact that I'm more reliable about the contents of my beliefs than any other person, not that I'm more reliable than other people in general.

Second and deeper, I am *exactly* as reliable as other people in general. Knowledge of the contents of one's own beliefs can be represented as homophonic T-sentences for sentences with those contents. But those T-sentences make reference to the truth-conditions of the embedded sentences, and those truth-conditions are determined, in triangulation, to be the contents of the embedded sentences. Knowledge of the contents of one's attitudes is reliable because it perfectly tracks the standard. No particular interpreter constitutes that standard, so any particular interpreter is less reliable than the believer herself. Nevertheless, the body of interpreters as a whole and across time are no less reliable than I am.

3 Scepticism and Relativism

3.0 INTRODUCTION

In this chapter, I begin to apply the theses developed in the previous chapters to issues outside of metaphysics, philosophy of mind, and philosophy of language. Recall that my overall goal is to defend moral realism. My main foils are expressivism and moral scepticism. But moral scepticism comes in two kinds: as an implication of general scepticism, and as a distinctive doctrine distinguishing moral beliefs from others. This chapter refutes the first kind, by refuting scepticism in general, and prepares for the refutation of the second kind.

In the first section, I take what might seem to be a detour through Kantian metaphysics. However, this section is in fact crucial to my overall approach. In it, I discuss Kant's approaches to scepticism and relativism. Of a non-exegetical purpose, I explore some of Kant's arguments in philosophical psychology and a notion of the synthetic *a priori* to which I will appeal later in the work.

In the second section, I refute scepticism. Davidson has offered two arguments against scepticism at various times. One of them, the oft-discussed "omniscient interpreter" argument, fails to appeal to Davidson's core theses in philosophy of mind and language. It also comes up short, requiring facilitation by the very theses that were not employed in it. I offer an interpretation and defense of Davidson's more mature anti-sceptical argument.

In the third section, I turn to conceptual relativism. Davidson has remarked that this view is either unintelligible or boring, and I agree. I canvass some relativist approaches, and find them to be either benign and in no interesting way relativistic. Finally, I offer Davidson's positive argument against relativism. Anti-relativism is

important for the version of moral realism I lay out in chapter 5, as well as my response to certain moral sceptics.

3.1 KANT'S REFUTATIONS OF IDEALISMS

In this section, I take a brief Copernican break from the constant linguistic turnings of the rest of this work. I want to discuss three points. First, I want to relate the Davidsonian position that I'm developing to the Cartesian view criticized by Kant in the *Paralogisms*. One might be dismissive of Davidson's project for much the same reason Kant was critical of Descartes', but that would be a mistake. Second, I want to discuss the off-hand remarks about Berkeley's "dogmatic" idealism in the *Refutation of Idealism*. While Kant's allusive critique of Berkeley is of limited interest to me here, there is another way of looking at Berkeley (and Leibniz) that makes him parallel contemporary conceptual relativists, and another way of looking at Kant that makes him parallel Davidson's critique thereof. Third, I examine the critique of Descartes' *cogito* in the *Refutation of Idealism*, and show a close parallel between Kant's approach to Descartes and Davidson's approach to scepticism. By discussing these points, I hope to accomplish three ends. First, I lend some historical depth to my treatment (which otherwise might seem to fail to acknowledge any philosophy from before 1879). Second, I assimilate the entire project of this dissertation to a certain Kantian tradition that I find amenable to my goals in moral and political philosophy. Third, I introduce and anticipate the arguments of the rest of this chapter.

3.1.1 Rational Psychology

Consider Descartes' "rational psychology," as Kant calls it in the first chapter of the *Paralogisms*. Rational psychology is the attempt to deduce substantive truths about the self from the assertion that I think:

> Now *rational* psychology actually is an enterprise of this kind. For if the slightest empirical [element] of my thought — some particular perception of my inner state — were also mixed in with this science's bases of cognition, then it would no

> longer be rational but *empirical* psychology. Hence we are indeed facing an alleged science which has been built on the single proposition *I think*... (A342/B400)

Someone would be forgiven for thinking that Davidson's project is an updated version of rational psychology. Davidson begins with premises about language: that its leading features are graspable on the basic of publicly available evidence, that it is compositional in form, and so forth. On the basis of the premises about language, Davidson tries to deduce some quite remarkable conclusions: there is no diversity of conceptual schemes, we are in touch with an external world and other minds, there is no non-coincidental type-identity between any mental and any physical types, there can be no theory of truth.... The surprising move is from the subjective — the structure of consciously spoken language — to two areas of the objective, inner and outer. Davidson deduces conclusions about thinkers that extend well beyond facts about their subjective linguistic experience. Also, he deduces conclusions about the external world. Descartes, trapped within the demon's illusions, can't move from the subjective to both the inner and the outer, but he does move from the subjective to the inner objective. He moves from his own subjective thoughts to facts about a certain object, his own mind, that extend well beyond experience, such as the substantiality, simplicity, and immortality of the soul.

> What is the problem with rational psychology?
>
> ...we can lay at the basis of this science [rational psychology] nothing but the simple, and by itself quite empty, presentation *I*, of which we cannot even say that it is a concept, but only that it is a mere consciousness accompanying all concepts. Now through this *I* or *he* or *it* (the thing) that thinks, nothing more is presented than a transcendental subject of thoughts = *x*. (A345-6/B404)

The *cogito* presents either an empirical object of awareness, or a transcendental subject of awareness. If an empirical object, then rational psychology has been reduced to some sort of introspective empirical psychology. If a transcendental subject, then there is precious little to be said about it: it's transcendental, so no predicates apply to it that apply only to the realm of objects of awareness, such as substantiality and permanence. But these are

just the sorts of things Descartes wanted to show applied to the *I* of the *cogito*. By one of the standard tropes of the critical method, Kant has shown the problem with Descartes' rationalist approach to self-knowledge.

Does a similar criticism apply to Davidson? Davidson is not an empirical psychologist or linguist. His arguments about the nature of language or mind are not based in empirical observation. Yet he comes to substantive conclusions with empirical application. Isn't Davidson just another rationalist?

No. Rather, Davidson's approach mirrors not Descartes', but Kant's own. Like Kant, Davidson offers transcendental arguments; in his case, arguments from the possibility of language. A comparison to another Kantian, the economist Ludwig von Mises, may be helpful. Like Davidson, von Mises studies the structure of the human mind on the basis of the nature of human action. Where Davidson tends to focus on linguistic action, since his goal is to understand intentional attribution, von Mises broadens his focus to action as a whole, but also takes more or less for granted our ability to *interpret* others' actions. In discussing the discipline he calls 'praxeology,' effectively what we call 'decision theory,' Mises says:

> The scope of praxeology is the explication of the category of human action. All that is needed for the deduction of all praxeological theorems is knowledge of the essence of human action.... No special experience is needed in order to comprehend these theorems.... The only way to a cognition of these theorems is logical analysis of our inherent knowledge of the category of human action. We must bethink ourselves and reflect upon the structure of human action. Like logic and mathematics, praxeological knowledge is within us.... All the concepts and theorems of praxeology are implied in the concept of human action. (Mises 1949/1966, p. 64)

Is it not plainly a mistake, on a par with rational psychology, to say that all of decision theory or economics is an *analytic* consequence of one *a priori* truth? Yes and no. von Mises's self-interpretation would make his method mistaken like rational psychology, but a better interpretation of his approach would make him Kantian. The point becomes both murkier and clearer here:

> [Praxeology's] statements and propositions are not derived from experience. They are, like those of logic and mathematics, a priori. They are not subject to verification or falsification on the ground of experience and facts. They are both logically and temporally antecedent to any comprehension of historical facts. They are a necessary requirement of any intellectual grasp of historical events. Without them we should not be able to see in the course of events anything else than kaleidoscopic change and chaotic muddle. (*ibid.*, p. 32)

Mises is even clearer here that praxeology is intended to be analytic *a priori*. But on the other hand, the theses of decision theory allow us to interpret experience. When we observe human action, we could not comprehend what we observe without applying to it what we know, *a priori*, about human action, and decision theory is nothing but a formalization of that *a priori* knowledge.

But this shows the same structure as Kant's positive contribution in the Transcendental Aesthetic and Transcendental Analytic, not a structure that he subjects to critique in the Transcendental Dialectic. Economic categories are necessary for the interpretation of experience of a certain kind, and they automatically hold of the relevant sort of experiences. They are, in Kant's terms, Pure Concepts of Understanding. Mises is plainly wrong, then, when he says that "Aprioristic reasoning is purely conceptual and deductive. It cannot produce anything else but tautologies and analytic judgments." (*ibid*, p. 38) As Barry Smith notes:

> Austrian economics seems to be like other *a priori* disciplines in that it involves a multiplicity of concepts connected together not hierarchically but rather in a dense holistic network of mutual connections whose order is not capable of being antecedently established. ...in Mises, we are dealing with a family of *a priori* categories and categorial structures which are — in contradistinction to Mises's self-interpretation but still in concordance with his actual practice in economics — not analytic but synthetic. (Smith, Barry, 1994. p. 316)

What we have in Mises and Davidson are two interpretive schemes, praxeology and Tarskian truth theories, known and knowable *a priori*[45], such that it's constitutive of human action or speech that these phenomena be interpretable according to the schemes.

---

[45] No particular Tarskian theory for a speaker is known *a priori*, of course, since we have to observe the speaker and create the theory. But the general form of the theory is *a priori* in the same sense as Kant's categories.

Davidsonian reflections on language don't try to tease out the implications of some opening definition. Indeed, Davidson argues that there *can be no* opening definition in the theory of meaning: the opening concept, truth, is too fundamental to be defined and so there can be no analysis of it. Trying to offer such a definition would be rational psychology; rejecting the attempt is Kantianism of the highest order. Rather, Davidson, armed with Tarskian theory, gives us a transcendental perspective on language and thought. My argument in section 2.3 that a language must be compositional is in line with this approach. I didn't argue analytically on the basis of a definition of language that language must be compositional; nor did I observe that all actual languages were compositional. Rather, I argued synthetically, on the basis of the pre-conditions of linguistic comprehension. I offered what we can call a transcendental deduction of compositionality. Kant's critique of rational psychology, then, can no more apply to Davidson's frankly Kantian approach than it does to Kant's own Transcendental Analytic.

3.1.2 Dogmatic Idealism

I want to move on to the *Refutation of Idealism* and Berkeley. My first goal here is to draw an analogy between an idealism that one can find in Berkeley and Leibniz and contemporary conceptual relativism. Then, I want to try to draw some lessons about how we should respond to conceptual relativism. On Berkeley's idealism, Kant says:

> ...the *dogmatic* idealism of Berkeley... declares space, with all the things to which space attaches as inseparable condition, to be something that is in itself impossible, and hence also declares the things in space to be mere imaginings. ...the basis for this idealism has already been removed by us in the Transcendental Aesthetic. (B274)

Berkeley argues[46] that we have no concrete idea of space; rather, we have the sensation of the power of motion and the expectation of tactile sensations on the basis of visual sensations and *vice versa*. As these sensations define the space they occupy, there is no space distinct from the objects that constitute it; an empty space is no space at all.

[46] Berkeley 1710, p. 113

Kant rejects this notion of relative space. There is only one context in which it makes sense to talk of things spatially: the intuition of physical objects. But that we experience them in space is a pre-condition of our experiencing these objects. It is thus vain to suggest that space supervenes on objects: on the contrary, the very possibility of the objects supervenes on the space in which they're perceived.[47]

But I want to take a different angle on this encounter. Consider how things must be to two people in a Berkeleyan universe. Each of them experiences ideas, including the sense of freedom of motion, and each of them anticipates tactile contacts on the basis of visual experience and *vice versa*. But, as they share no ideas, and space is identified with reference to its contents, they share no space. This *privacy* of space seems to be a consequence of Berkeley's *relativity* about space.

The same idea appears in Leibniz's *Discourse*. Leibniz says:

> ...the ideas of size, figure and motion are not so distinctive as is imagined, and... they stand for something imaginary relative to our perceptions as do, although to a greater extent, the ideas of color, heat, and the other similar qualities in regard to which we may doubt whether they are actually to be found in the nature of the things outside of us. (Leibniz 1686/1902, §XII, p. 18)

Further:

> ...every substance is like an entire world and like a mirror of God, or indeed of the whole world which it portrays, each one in its own fashion.... Thus the universe is multiplied in some sort as many times as there are substances... (*ibid*, §IX, p. 15)

From the relativity of space, Leibniz adopts the privacy of space. Space is literally within the subject, and each subject has her own internal space with no contact with other such spaces.

For Leibniz and (I suggest) Berkeley, space is private and idiosyncratic to each subject. That we can never share objects of awareness is either a premise supporting the conclusion of the privacy of space, or a consequence drawn from that alleged privacy. There is an analogy between the relativity of space to subject in early modern philosophy,

[47] Kant 1787/1986, A22/B37-A31/B45

and the relativity of conceptual scheme to subject in contemporary philosophy. For Berkeley space was defined with reference to the objects that appeared in it. But the objects were private. Hence space was private. For Leibniz, space was defined with reference to the objects contained in it. But the space was private. Hence the objects were private. In contemporary philosophy, figures like Kuhn have argued that the objects of awareness are defined with reference to the conceptual scheme employed to grasp those objects. Since the conceptual scheme is private, so are the objects.

Leibniz and conceptual relativists have similarly structured arguments, and the position of Berkeley shows a distinct structural similarity to Leibniz's. So we might gain some insight into how to defeat conceptual relativists by looking at how Kant might have argued against Berkeley or Leibniz. On the notion of multiple (including multiple private) spaces, Kant says:

> ...we can present only one space; and when we speak of many spaces, we mean by that only parts of one and the same unique space. Nor... can these parts precede the one all-encompassing space, as its constituents, as it were (from which it can be assembled); rather, they can only be thought as *in it*. Space is essentially one; the manifold in it, and hence also the universal concept of spaces as such, rests solely on [our bringing in] limitations. (A25/B39)

It's a mistake to think that there could be more than one space. If we reflect on our attempt to give concrete meaning to the notion of multiple spaces, we'll find that we relate the spaces to one another, spatially. The only space is the space in which we find these "multiple" spaces. But if there can be only one space, then, if space is idiosyncratic to a subject, then it is idiosyncratic to *that* subject, and no other subject can experience objects in space. So Berkeley and Leibniz must either give up the privacy of space, or else deny the existence of more than one subject with spatial experience.

The analogy with conceptual relativism is clear. Conceptual relativism identifies a conceptual scheme by reference to the objects that can be grasped in it, but also identifies those objects by reference to the schemes in which they can be grasped. It proceeds to have it that the schemes are private or at least multiple. To show that conceptual

relativism is false, then, it must be shown that, like space, conceptual schemes are exactly one in number. If I could show all conceptual schemes had to be conceptually related to one another, just as all private spaces have to be spatially related to one another, perhaps I would have solved the problem.

Likewise, the early modern example constitutes a cautionary tale. Kant's argument for the unity of space rests on his dogmatic Euclideanism. If my argument against conceptual relativism is to share some sort of general structure with Kant's argument and yet be successful, it will need to avoid making a similar error. I would make a similar mistake if I were to identify some cluster of concepts or judgments and assert that these particular items must appear in every conceptual scheme, and having those items in common is sufficient for identity of conceptual scheme(s). Whatever items I chose, I would have done so on more or less the same basis that Kant asserted that space is necessarily Euclidean: I'm not imaginative enough to conceive of a conceptual scheme that lacks them. Such a dogmatic approach would undermine an anti-relativistic argument.

3.1.3 Sceptical Idealism

I want to move on to the rest of the *Refutation of Idealism* and its attack on Descartes' scepticism: the position that Descartes would have had, had he not indulged in the Cartesian circle in the *Third Meditation*. Kant's argument will show structural similarity to Davidson's. Each of them will contend that the self, Descartes' *cogito*, can only be understood or identified against the background of an external world that is largely as one understands it to be. Kant argues as follows:

> Theorem
>
> *The mere, but empirically determined, consciousness of my own existence proves the existence of objects in space outside me.*

Proof

> I am conscious of my existence as determined in time. All time determination presupposes something *permanent* in perception. But this permanent something cannot be something within me, precisely because my existence can be determined in time only by this permanent something. Therefore perception of this permanent something is possible only through a *thing* outside me and not through mere *presentation* of a thing outside me. Hence determination of my existence in time is possible only through the existence of actual things that I perceive outside me. Now consciousness of my existence in time is necessarily linked with consciousness of the possibility of things outside me, as condition of the time determination. I.e., the consciousness of my own existence is simultaneously a direct consciousness of the existence of other things outside me. (B275-6)

The leading idea of Kant's argument is that there is no self-awareness without awareness of a world beyond the self. To experience myself as an object continuing in time, I have to have some sort of standard or background against which to attribute myself continuing existence in time. This contention leads to three questions. First, why should there be such objects at all? Kant agues:

> ...time by itself cannot be perceived. Hence the substrate which presents time as such, and in which all variation or simultaneity can in apprehension be perceived through the appearances' relation to it, must be found in the objects of perception, i.e., in the appearances. (A182/B225)

All experience, Kant contends, is in time; moreover, in a single line of time. But to experience something in time, it must be experienced against some relatively enduring background. If it weren't, then there would be no sense in thinking of time as a single line; rather, we would experience time as, as it were, ending and beginning anew, which contradicts the notion of time's being a single line.

Second, why must there be only one such object? For much the same reason. To whatever degree the background is broken up, our sense of a continuous time is also broken up. However, this line of argument does not imply that there must be some specific object, like the Earth or one's body, that one must use as background. Kant's claim in the *First Analogy* is that "*In all variation by appearances substance is permanent, and its quantum in nature is neither increased nor decreased.*" (A182/B224)

The *Analogy* does not claim that any particular substance, in the sense of physical object, is permanent; nor must we treat the *Analogy* as trying to prove the conservation of matter *a priori*.[48] Strawson makes this suggestion:

> Kant, let us say, has shown the necessity of something abiding and permanent, viz. the whole frame of Nature; and if the word "substance" is to be linked with the concept of absolute permanence, it is to the whole frame of Nature that it should be applied — as it was by Spinoza. (Strawson 1966, p. 130)

If Strawson is right, then the point of the *First Analogy* is that, in order to identify anything as existing in time, we must identify it against the background of the world as a whole.

Third, why can't *I*, rather than the external world, form the background? This question returns us to the *Paralogisms* and the attack on rational psychology. Someone can use 'I' in two senses. In one sense, 'I' is the subject of all experience, which is not an empirical object of experience. But then no empirical categories, like time, can be applied to it. So this "object" can't be the background against which I identify events in time. In the other sense, 'I' refers to one among many objects of experience, the self and its beliefs, desires, memories, experiences, and so forth. But:

> ...this permanent something cannot be an intuition within me. For all bases determining my existence that can be encountered within me are presentations; and, being presentations, they themselves require something permanent distinct from them, by reference to which their variation, and hence my existence in the time in which they vary, can be determined. (*ibid.*, p. 290 note 46)

The empirical referent of 'I' is, as Hume said, a fleeting bundle of impressions, not the ultimate background against which we identify ourselves as continuing across time. The empirical self is one of the objects that can be identified in time only relative to the background of something permanent. The point here is that the self can itself only be experienced coherently against the background of the world. In the absence of veridical worldly experience, I couldn't identify myself. Descartes is wrong. We need not argue our way out from subjective experience to an objective world. The objective world is

[48] See Allison 1983, pp. 199-215

what we experience, and the self *exists* only as a consequence of its experience of the objective world: "...inner experience as such is possible only through outer experience as such..." (B278-9)

Davidson's argument against scepticism is strikingly similar. For Davidson, that a thought exists and has content is a consequence of its place in its causal and logical context. Without the external world and other speakers and interpreters, there is no thought; indeed, without being embedded in a context of mostly true beliefs, nothing can count as a thought. Given these similarities, having a grasp on Kant's refutation of scepticism should help lead to a sharper formulation of Davidson's refutation of scepticism. At a minimum, we can see a precedent for Davidson's approach in the work of a philosopher no one can dismiss.

3.2 SCEPTICISM

3.2.1 Preliminaries: The Sceptical Target and Transcendental Arguments

Here, I take it, is a standard presentation of the sceptical problem. Allow that ϕ and ψ are not, intuitively, compossible. ϕ is some sentence that I take to be true in virtue of the appearance of the external world immediately surrounding me; for instance, that there is a cat on the mat. ψ, on the other hand, is some outlandish sceptical hypothesis, such as that I am the victim of (pointlessly) deceptive artificial intelligences who (pointlessly) use my body heat to supplement their nuclear fusion to power themselves, and who have arranged things so that my experiences and beliefs entirely fail to represent the world around me: where I think there is a cat on the mat, in truth, there is no cat and no mat but just a carapace designed to house my physical body and (pointlessly) collect heat from it. Because ϕ and ψ are not compossible, it seems that, if I believe that one of them is true and I consider the other possibility and their incompatibility, then I should be able to deduce that the other is false; alternatively, if I cannot convince myself that the other is false, then I should be forced to withdraw my

belief that the first one is true. I should like to believe that ϕ, but, sadly, I find myself considering the possibility that ψ, as well as the incompatibility between ϕ and ψ, and discovering that I cannot convince myself that ψ is false. I am thus forced to withdraw my belief that ϕ is true. But I see that I can repeat this experiment for any other ϕ, so long as it's a sentence that I take to be true on the basis of the appearance of the world around me. So I'm forced, by the possibility of deceptive AIs (and my commitment to logic) to admit that I shouldn't really believe anything on the basis of the appearance of the external world around me.

Further light can be shed on the sceptical problem by considering a less extreme sceptical situation. I opine that ϕ, and a friend explains that I'm mistaken. There's no cat on the mat: there's a cat near the mat and a mirror on the mat. I check again and see the frame, get up and walk around the mirror, see where the cat is actually located, and confess that ϕ is false. It's worth wondering how this routine of checking the world could help.

The problem with ϕ and ψ is this. If the world is as I take it to be, ϕ is true in virtue of some circumstance obtaining in the world around me, and that very circumstance plays a crucial role in bringing it about that I believe that ϕ. The reason ϕ and ψ are not compossible is that if ψ is true, then everything that brought it about that I believed that ϕ has nothing to do with ϕ's being true, but rather to do with the AIs and their (pointlessly) dastardly plans. But from the inside, I can't tell whether ϕ or ψ is what brought it about that I believe that ϕ. I am, as it were, on one end of a line, with my belief that ϕ. On the other end of the line is either ϕ or ψ, helping to bring it about that I believe that ϕ. In one case, my belief is true; in the other case, it's false. But the only way that I could check is by *stepping off of the line* and looking at the trigger of my belief *from a different angle*. That's what I do when I get up to check about the mirror, but I can't do it with a sufficiently pervasive deception, such as that encompassed by ψ.

Since I can't stand beside the line that connects me with (or divides me from) the world, I need help. I need some third person to stand beside the line between my belief and its trigger, and guarantee that they have the right match. Descartes appealed to God to do that job, but it didn't work out. A more successful approach would have it that the existence of an attitude that's susceptible to being true or false relies on the existence of a third person, whose securing of my content is also sufficient to guarantee its truth. By building the third-person perspective into the existence and nature of content, we can guarantee that our beliefs are, for the most part, true.

Davidson's response to scepticism is, arguably, a transcendental argument, so I should consider how a transcendental argument is supposed to go. I begin with A.C. Genova's discussion of good transcendental arguments, which are to satisfy three criteria; such an argument:

> demonstrates transcendental principles (a unique and invariable conceptual core) as a necessary presupposition of *all* possible human experience with respect to a specifiable domain of objects,
>
> establishes the objectivity thesis, i.e., bridges any meaningful gap between the subjective necessity of the conceptual scheme and the specified objects..., and
>
> does this validly without the taint of verificationism. (Genova 1984, p. 476)

Since no one likes verificationism, any argument should do whatever it does without the taint of verificationism. (3), then, follows from the goodness of good transcendental arguments, not necessarily from their transcendence. But it's not immediately obvious what the relationship is between (1) and (2), and why both should be necessary for an argument both "good" and "transcendental."

It's possible to mistake criterion (2) for the criterion that a transcendental argument must defeat scepticism. That is not correct. An argument that satisfies (2) will show that there are objects answering to some specific conceptual scheme, not that most of our beliefs are true. The specific conceptual scheme in question is the one established by the argument if it satisfies criterion (1).

For Genova, a good transcendental argument will proceed in two phases, one to match criterion (1), and the other to match criterion (2). The first phase is a *metaphysical deduction*, an argument "which provides an a priori justification of a unique [conceptual framework] — a [conceptual framework] which is a necessary presupposition of all possible contingent interpretations of experience." (Genova 1984, 479-80) The second phase is a *transcendental deduction*, an argument "which provides an a priori justification for the objective validity of a unique [conceptual framework]..." (*ibid*, p. 480) The metaphysical deduction will select some conceptual framework as constitutive of an area of experience, while the transcendental deduction will show that the framework has objective application; that is, it applies to real objects. Genova will offer an interpretation of Davidson (to be discussed in 3.2.3) that treats his anti-sceptical argument as a good transcendental argument of this form.

Whether Davidson's anti-sceptical argument is a good transcendental argument by Genova's criteria or not (it is not), it is at least in the general vicinity of a transcendental argument. It's important, then, to be clear on a way in which Davidson's view is completely unlike Kant's. Maker offers a misleading or incorrect assessment of the nature of transcendental arguments, at least if Davidson can he held to offer any transcendental arguments:

> In both Kant and Davidson, the transcendental procedure involves two stages. 1. Objectivity is subjectivized or interiorized. 2. It is argued that the distinction between objectivity *as subjectivized* (how things appear [for Kant], how we describe them in language [for Davidson]) and some *radically other* objectivity (some conception of objectivity which might be inaccessible from the domain of subjectivity) is unintelligible or incoherent. (Maker 1991, pp. 349-50)

This has the appearance of being a bizarre misinterpretation. Whether it is misleading or incorrect, though, depends on just what Maker means.

As Kant interpretation, the description seems correct. Kant "interiorizes" objects by turning them into phenomena, which exist only in relation to the synthetic activity of a mind. It makes sense to say that, for Kant, objectivity as subjectivized *is* how things

appear. But as Davidson interpretation, it seems off. Objects, and how they are described in language, are not the same. Objects can exist quite independently of our describing them in language. What Maker has in mind with his strange phrasings, though, might be a bit more on the mark:

> In the case of Davidson, objectivity is seen as accessible in that our access to it is construed in terms of an operation present within subjectivity.... Davidson asks us to understand objectivity as necessarily immanent within language as understandable. ...objectivity is subjectivized in the sense that a procedure said to be inseparable from subjectivity — ...the determination of linguistic meaning — is explained as an activity which can be accounted for on only the condition of the involvement of objectivity in it. (*ibid*, p. 352)

Davidson's argument will certainly have it that thoughts can only exist against the background of the external world. If the essence of the transcendental approach is argument to the effect that the subjective can only exist against the background of the objective, then Davidson's approach is certainly transcendental. But Davidson's transcendence would be quite different from Kant's, since Kant's "interiorization of objectivity" is completely unlike Davidson's.

Barry Stroud notes that, "If [Davidson's claim that] 'belief is in its nature veridical' were true..., some comparable [to transcendental idealism] explanation would surely be needed of how and why such a remarkable thing must be true." (Stroud 1999, p. 158) Stroud is right, and the explanation has to do with Davidson's theory of interpretation, especially its externalistic features. So Davidson does have a scheme, as powerful in its way as transcendental idealism, to buttress an anti-sceptical argument. In chapters 1 and 2, I hope to have developed it adequately for present purposes. Unfortunately, the deployment of that scheme will have to await section 3.2.3. Section 3.2.2 will, for completeness' sake, take a detour through a defective presentation of the anti-sceptical argument.

3.2.2 The Omniscient Interpreter

Davidson's (in)famous omniscient interpreter argument has been rejected by no less an authority than Donald Davidson:

> ...the argument that summons up an Omniscient Interpreter does not advance my case. As with Swampman, I regret these forays into science fiction and what a number of critics have taken to be theology. If the case can be made with an omniscient interpreter, it can be made without, and better. (Davidson 1999a, p. 192)

An argument both science fictional and theological — a Lewisian argument, one might say — is surely a sad thing, as it strains credulity in at least two dimensions. If even its creator has abandoned it, it might not seem worthwhile to discuss it. For some philosophical purposes, this orphan could well be left to languish. However, there are also matters of interpretation and significance. In section 1.3, I discussed Davidson's two views on the nature of truth-conditions: according to one view, truth-conditions are theoretical constructs; according to the other, some of them are events in the world. In section 2.2, I mentioned alternative principles of charity. Some versions of the principle of charity sought to maximize truth in the speaker — which would beg the question against the sceptic — while some sought to maximize agreement between the interpreter and the speaker. As long as truth-conditions are theoretical constructs, the principle of charity that says to assign causes to utterances as their contents makes no sense: contents aren't real, so none of them are causes. But without the later principle of charity, Davidson's anti-sceptical argument can't work: it must take a form like the omniscient interpreter argument, which fails. The point of this discussion is to see exactly why the omniscient interpreter argument fails, and thereby show the significance, to an area of philosophy outside philosophy of language, of the choice we face in selecting a principle of charity.

Davidson presents the relevant notion of charity for the omniscient interpreter argument here:

> ...we damage the intelligibility of our readings of the utterances of others when our method of reading puts others into what we take to be broad error. We can make sense of differences all right, but only against a background of shared belief. What is shared does not in general call for comment: it is too dull, trite, or familiar to stand notice. But without a vast common ground, there is no place for disputants to have their quarrel. (Davidson 1977a, p. 200)

The principle of charity Davidson employs in the omniscient interpreter argument is the principle of massive agreement. This is actually an attractive principle, until it's compared with the principle of assigning causes as contents.

Consider the discussion of holism from section 2.3. If I fail to assign the appropriate broad swath of beliefs, then I fail to interpret at all. But the appropriate broad swath must find a match in my own beliefs. Interpretation proceeds through the formation of a Tarskian truth theory for the speaker's utterances. Such a theory requires a match between the speaker's utterances and those of my utterances that I use to interpret her. If I systematically assign false (by my lights) sentences of mine as interpretations of the speaker's utterances, then it's not plain that I'm interpreting her. If she doesn't have *any* belief that I have about some sort of object, then I violate the intensionality of attitude attributions by assigning her any beliefs at all about that sort of object. I cannot, for instance, assign the speaker any beliefs at all about rabbits if she has *no* belief that I can identify as helping to fix, by logical implication, *rabbits* as relevant to any her beliefs. You can't believe that *rabbits* are reptiles unless you also believe that rabbits are fuzzy, eat carrots, tend to procreate at a rapid rate, and so forth in an undefined mass of beliefs that I hold. Denying that you hold *any* belief that I hold about rabbits makes it impossible for me to assign you *any* belief about rabbits, whether I share it or not.

The fact that this argument flows — we do have to massively agree with speakers in order to interpret them — does not imply that finding agreement is the essence of charity. But Davidson latched on to this feature of charity and used it as the basis for the omniscient interpreter argument.

I must find the speaker to be massively right, by my lights. So, of course, when I take up the position of speaker, anyone who interprets me must find me to be massively right, by her lights. But why can't we be wrong together?

> We do not need to be omniscient to interpret, but there is nothing absurd in the idea of an omniscient interpreter; he attributes beliefs to others, and interprets their speech on the basis of his own beliefs, just as the rest of us do. Since he does this as the rest of us do, he perforce finds as much agreement as is needed to make sense of his attributions and interpretations; and in this case, of course, what is agreed is by hypothesis true. But now it is plain why massive error about the world is simply unintelligible, for to suppose it intelligible is to suppose there could be an interpreter (the omniscient one) who correctly interpreted someone else as being massively mistaken, and this we have shown to be impossible. (*ibid*, p. 201)

I take it that this argument may be fairly represented as having three premises:

> (1) For any interpreter and interpretee, the interpreter *must* accept most[49] of the interpretee's beliefs;
>
> (2) It *must* be that every belief held by the omniscient interpreter is true;
>
> (3) The omniscient interpreter interprets me.
>
> Thus, Most of my beliefs are true.

The argument appears to be valid. So if there is a problem, then it is in the premises. (1) is supposed to be demonstrated on the basis of the holism and intensionality of attitude attributions. (2) is obviously true in virtue of the meaning of 'omniscient.'

(3) seems to be false, since nothing is omniscient; or, in any event, (3) is false so far as we know. That fact inspired the first formal critique of the omniscient interpreter argument. Foley and Fumerton argue that:

> From [my premise (1)] If there were an omniscient interpreter employing Davidson's methods of interpretation he would believe that most of what Jones

[49] "Most" is not a standard quantifier. The obvious fallacy that someone could fall into by using claims about "Most Fs" is transitivity. But I have only one "most" quantifier in the premises, so I don't make that error. Davidson (Davidson 1983, pp. 138-9) says, puzzlingly, that "...there is no useful way to count beliefs, and so no clear meaning to the idea that most of a person's beliefs are true." He then recasts the claim that most of our beliefs are true as the claim that our actual beliefs get a strong presumption of truth. But it's not clear why there's no way to count beliefs. It can't be that there are infinitely many beliefs, since, if there were infinitely many beliefs, then one could believe the infinitely many theorems of a Tarskian truth theory, which one (Davidson contends) can't.

> believes is true, how is it supposed to follow that most of what Jones believes is true? From [(1)] we *can* infer [(3) → (Conclusion):] If there were an omniscient interpreter of Jones employing Davidson's methods, most of what Jones believes would be true. But surely we need to affirm the antecedent of this conditional if we are to conclude that most of Jones's beliefs *are* true. That is, we need to affirm that there *is* an omniscient interpreter of Jones. (Foley and Fumerton 1985, p. 84)

All we can actually get, Foley and Fumerton point out, is that *if* (3) *were* true (which it isn't), *then* most of my beliefs are true.

But Davidson didn't exactly say (3). What Davidson said was that "there is nothing absurd in the idea of an omniscient interpreter." One thus gets the urge to reform (3) along these lines: (3*) Possibly, an omniscient interpreter interprets me. But consider the merely possible existence of the omniscient interpreter. The problem that Foley and Fumerton point out is that *possibly* being omniscient and interpreting me doesn't make the omniscient interpreter agree with, and therefore guarantee, my *actual* beliefs. Maybe, since I'm massively wrong, the only way to actualize the possibility that I'm interpretable by the omniscient interpreter would be for me to totally change my beliefs. Brueckner replies that omniscience doesn't stop at the edge of reality:

> Davidson need make no assumptions concerning possible worlds containing both me and an OI [omniscient interpreter]. Instead, he just needs this assumption:
>
> > (A) Some possible world W* contains an OI who has perfect knowledge about all possible worlds, including the actual world; thus he believes, among other things, all and only true propositions about the actual world.
>
> In other words, there might have been an OI with perfect knowledge about all possible worlds; if such a being had existed, he would have believed, among other things, all and only true propositions about this, the actual world. For all that has been assumed in (A), W* is not the actual world and does not contain me. (Brueckner 1991, p. 201)

What Brueckner has in mind is that the merely possible omniscient interpreter would still agree with me about my world, were he to engage in trans-world interpretation. So possible agreement would be sufficient for agreement between the omniscient interpreter and me; even if she is not actualized in the actual world, her interpreting me is. So we may reform (1):

(1*) For any *possible* interpreter and interpretee, the interpreter *must* accept most of the interpretee's beliefs;

The argument obviously flows through despite these alterations.

But (3*) makes no sense on Brueckner's version of the argument. His version of the argument relies on modal realism: for if there weren't *really* other possible worlds for the omniscient interpreter to live in, then it would be pointless to discuss the omniscient interpreter's interpretation of me. We would be discussing a fantasy *as such* — not the same as a thought experiment, in which we discuss a fantasy *as though it were real*. But the possibility that's to be realized is a trans-world possibility: that the (real but non-actual) omniscient interpreter interprets me across worlds. If this bizarre claim is going to appear as a premise, then whether it is possible or not should be a subject for discussion, but, on the possible worlds semantics for modal talk, it isn't: possibility is reality in some possible world, but no relation *between* possible worlds is real *in* any world at all. Further, we're talking about what *would* be the case *if* the omniscient interpreter *were to* interpret us from afar: this is counterfactual talk, which we again can't make sense of when the counterfactual relates things from different worlds. Modal realism, especially in this context of trans-world relations, brings in train unintelligible talk about trans-world possibilities and counterfactuals, so any argument that relies on possible worlds realism has a corrupted approach to its modal aspects.[50]

Brueckner's approach doesn't seem to save the argument from the Foley and Fumerton objection. There is an additional premise, however, that would solve the problem. The sceptic requires that I have beliefs; else, I have nothing to be wrong about. But what if interpretability by the omniscient interpreter were a prerequisite on having beliefs?

'Interpretability' is a modal notion, and modal talk in this area is getting vexed. Here's what I mean, in the possible worlds jargon. x is interpretable by y iff, in some

[50] Thanks to Neil Sinhababu for helpful discussion on these modal matters.

world in which x has the same beliefs, with the same truth values, that he has in the actual world, y interprets x. y need not exist in the actual world, of course, as the omniscient interpreter does not. The premise, then, is that, if someone has beliefs, then, in some world in which she has just those beliefs with just the same truth-values, the omniscient interpreter interprets her. By tying my beliefs in the world in which I am interpreted to my beliefs in this world, this premise would defeat the Foley and Fumerton objection construed as an objection to the validity of the argument.

The argument would need reform elsewhere, of course. For instance, we would need the premise that I do, in fact, have beliefs. The new premise set, then, is:

> (1*) For any *possible* interpreter and interpretee, the interpreter *must* accept most of the interpretee's beliefs;
>
> (2) It *must* be that every belief held by the omniscient interpreter is true;
>
> (3**) If anything interprets me, then it *must* be that an omniscient interpreter *could* interpret me;
>
> (4) Something interprets me;
>
> Thus, Most of my beliefs are true.

This argument (modulo the steps I skip) also seems to be valid. And again, except for premise (3**), it seems sound. But what are we to make of (3**)?

Allowing 'n' to refer to a noniscient interpreter, an interpreter absolutely wrong about everything, consider replacing (2) and (3**) with (2') and (3'):

> (2') It *must* be that every belief held by the noniscient interpreter is false;
>
> (3') If anything interprets me, then it *must* be that a noniscient interpreter *could* interpret me.

From this premise set, one can conclude that most of one's beliefs are not true. Why should interpretability by the omniscient, rather than the noniscient, interpreter, be a prerequisite on having beliefs?

What exactly is the problem with the noniscient interpreter? Of course, if the original argument succeeded, then there is no such thing as a noniscient interpreter, since

interpretation requires beliefs and beliefs are, by nature, massively true. But, on the other hand, if this second argument succeeded, then there could be no such thing as an omniscient interpreter, since interpretation requires beliefs and beliefs are, by nature, massively false.[51]

If interpretability by an omniscient interpreter is made a prerequisite on having beliefs, then we can discover that any believer must be mainly right. If interpretability by a noniscient interpreter is made a prerequisite on having beliefs, then we can discover that any believer must be mainly wrong. But it's not plain at this point why interpretability by *anyone* has anything to do with beliefs. The omniscient interpreter argument proceeds as though the issues of triangulation and externalism had never emerged. Why can't I have beliefs without being interpretable, by ordinary thinkers or thinkers extraordinarily gifted or inept? If I can be a believer without being interpretable, though *were* it the case that, were I interpreted, I would be mainly right, nevertheless, since I can't be interpreted, I can still be mainly wrong.

In offering the omniscient interpreter argument, Davidson focused on the form of a Tarskian truth theory to the exclusion of the procedure of radical interpretation. It's true that an interpreter must massively agree with the speaker in order to interpreter her. But the additional premise that's needed is that there must be an interpreter for there to be speech and belief at all. The new premise (3**) leads in this direction, but once we attend to the externalistic justification for the premise, we can skip that "premise" and directly conclude that scepticism is false. The omniscient interpreter argument is best understood as Davidson's anti-sceptical argument, shorn of its roots in radical interpretation, externalism, and the social nature of language. I now turn to the actual argument.

---

[51] To the best of my knowledge, this reply to Davidson was first offered in Dalmiya, 1990. It has also appeared in Carpenter, 1998, Ludwig 1992, and Ludwig and Lepore 2005 (pp. 326-9).

3.2.3 The Nature of Content

I want to introduce the argument by discussing two interpretations of it, Peter Klein's and A.C. Genova's. Klein will correctly interpret the argument, but incorrectly assess its soundness; Genova will incorrectly interpret the argument, but correctly assess its soundness. The failure of Genova's interpretation will help us see exactly the sense in which Davidson's argument is not a (Genovan) good transcendental argument, and exactly what pre-conditions on the existence of contentful attitudes are relevant to the argument. It should emerge from the discussion that the argument is sound and that scepticism is false.

As I understand it, the argument can be understood to proceed in two phases. The first phase uses externalistic premises to secure the truth of all normal observation beliefs (on a conception of normalcy for beliefs guaranteeing that the overwhelming majority of observation beliefs are normal). The second phase uses that conclusion, and content holism, to secure the truth of most non-observation beliefs. The conclusion is that most of our beliefs are true. One of the best general statements of the first phase of the argument appears here, in my favorite passage:

> What stands in the way of global skepticism of the senses is... the fact that we must, in the plainest and methodologically most basic cases, take the objects of a belief to be the causes of that belief. And what we, as interpreters, must take them to be is what they in fact are. Communication begins where causes converge: your utterance means what mine does if belief in its truth is systematically caused by the same events and objects. (Davidson 1983, p. 151)

In radical interpretation, I formulate laws in the form of Tarskian theorems that relate a speaker's utterances to the situations that prompt those utterances. The utterances have those situations as their *meanings*. In virtue of meaning those situations, the utterances are connected with external reality in a way that defeats the sceptic.

Klein contends that Davidson's argument is circular or groundless. Here is a requirement on any argument designed to defeat the global sceptic: "*...an argument against global skepticism cannot employ any premise whose plausibility depends upon*

knowledge of the actual world. To do so would clearly presuppose that we have some of the knowledge which the skeptic questions." (Klein 1986, p. 372) If we take for granted that we have knowledge of causes of beliefs, then we take for granted that we have knowledge of the external world, which would make the argument circular. But if we fail to take that for granted, then the argument from content is groundless. The essential feature of this reading is that it makes the causal determination of content the ultimate ground of the argument.

Genova argues that this reading of Davidson is superficial:

> The [anti-sceptical argument][52]... will not depend merely on the methodological thesis that belief content cannot be specified independently of the causes of belief, or on the supportive premise that scheme-content dualism is untenable. ...the validity of the [anti-sceptical argument] depends upon a still more fundamental background assumption, viz., Davidson's primary conclusion in his 'On the Very Idea of a Conceptual Scheme': translatability is a criterion of languagehood. (Genova 1999, p. 176)

Notice the structure that Genova imposes on Davidson's argument. First, Genova says, Davidson will establish that there is only one (or, to put it another way, exactly zero) conceptual scheme(s). Then, and on that basis, he will establish that that conceptual scheme has objective validity. This structure is the structure of a Genovan good transcendental argument. First, we have a metaphysical deduction of a conceptual scheme as constitutive of experience of a certain category of objects. In this case, the conceptual scheme is our ordinary conceptual scheme, the one displayed in ordinary language and science. Second, we have a transcendental deduction of the objective validity of that conceptual scheme. In this case, the transcendental deduction is the argument against scepticism.

For Genova, Davidson's attack on conceptual schemes is essential to the anti-sceptical argument, and the causal determination of content is not. However, Genova's

[52] Genova calls Davidson's anti-sceptical argument the Omniscient Interpreter Argument, whether the version of the argument under discussion invokes an omniscient interpreter or not. Since that strikes me as misleading, I replace his abbreviation 'OIA' throughout.

interpretation of the opposition to conceptual schemes just *is* the causal determination of content:

> The idea of a system of beliefs or sentences which is not translatable or not interpretable is unintelligible. Consequently, if the idea of a coherent system of massively false beliefs is intelligible, it must be interpretable.... Interpretability, of course, is defined in terms of Davidson's methodology of interpretation; an interpreter constructs an empirical theory of truth for the speaker (as well as a theory of the speaker's beliefs, etc.) that conforms to the formal and empirical constraints on such a theory, and proceeds only on the basis of the relevant data for interpretation as grounded on the principle of charity — taking what one construes as the causes of beliefs as their truth conditions. (*ibid.*, p. 180)

For Genova, the fact that there is only one conceptual scheme is the fact that we identify causes of beliefs as their truth conditions. The contrast between Klein's and Genova's interpretations of Davidson's argument, then, loses its edge, and Genova's implicit claim that Davidson's argument is a good transcendental argument loses its point. The core of the argument is that interpretation relies on causal relations between utterances and their prompting circumstances.

Now let me turn in earnest to the argument itself. To begin with, I need to recall two technical notions I defined in 2.2: the belief set and truth-conditional predicate. Two beliefs belong to the same belief set just in case they have the same truth-condition. A truth-conditional predicate is some predicate satisfied by each of the members of a set of truth-conditions and nomically correlated with the existence of members of some belief set. With these definitions in place, I can state and defend the first premise of the anti-sceptical argument:

(1) For all normal observational beliefs of mine,

> (a) there is a belief set, of which it is a member, and that has a truth-conditional predicate; and
>
> (b) there is a satisfier of that truth-conditional predicate, and that satisfier caused the belief.

Why do I refer to myself in the premise? Like all anti-sceptical arguments, this one must be framed in the first person, for to assume any premise not knowable from the first

person is to beg the question against the sceptic. The self-reference is present as a reminder. I will be even more important in the matching premise of the second phase of the argument.

Why do I refer to normality? Here, I frame the premise as a defeasible or commonsense conditional.[53] An anti-sceptical argument that purported to show that every belief is true would, of course, be hopeless. What I will try to show is that all normal beliefs are true. That might not seem very significant, unless there is some known ratio of normal to abnormal beliefs. But I characterize normality of belief in such a way that the overwhelming majority of beliefs must be normal.

The defense of the first part of the premise is very easy. Every belief has a truth-condition, so for every belief, not just all normal observation beliefs, there is a belief set (possibly consisting of just the one belief). Observation beliefs are just the beliefs that are nomically correlated with the occurrence of their truth-conditions, so all observation beliefs, not just normal ones, are members of belief sets with truth-conditional predicates.

The defense of the second part of the premise is likewise easy at this late date. The second part of the premise is defended with reference to Grace, the principle summed up in the slogan that cause is content. For something to count as the content of an observation belief, it must satisfy a predicate nomically coordinated with the existence of beliefs with that observation belief's truth-conditions. In virtue of the coordination's being nomic, such a satisfier must usually appear as the cause of beliefs with that belief's truth-conditions; else, there would be no (defeasible, *ceteris paribus*) law connecting them. A "normal" observation belief is just one showing the typical causal ancestry for beliefs of its kind; abnormal beliefs will be those whose causes are of the wrong kind. By

[53] On which see the system of defeasible natural deduction offered in Bonevac, 2003, pp. 434-74. I appeal to no special feature of Bonevac's system, and I can't imagine any controversy to which my use of the defeasible conditional could give rise; rather, if *my* use of the commonsense conditional were problematic, then *any* use would be problematic.

virtue of Grace, and hence the causally and socially externalistic basis of content, all normal observation beliefs are caused by their own truth-conditions.

> (2) Necessarily, for all observational beliefs of mine, truth-conditional sets of which they are members, and truth-conditional predicates of those sets, if there is an event which satisfies the truth-conditional predicate and caused the belief, then the belief is true.

The motivation (as distinct from the justification) of the necessity operator is that the argument is to show a defeasible conditional. From within the proof of a defeasible conditional, only modally closed sentences and some defeasible conditionals are available. Since I want the premise to be available within the defeasible conditional subproof, it must be modally closed. The justification is that the claim is a conceptual truth, hence necessary.

Each observation belief of mine is a member of some belief set, and all normal members of that set are caused by a satisfier of the truth-conditional predicate of the set. Here, the claim is that, if a particular member of a belief set is caused by such a satisfier, then the member is true. The proof of this premise again recurs to the identification of cause with content. What it is to have a certain content is to be caused by it (*modulo* the intricacies discussed in 2.2), but contents are truth-conditions. So what it is to have a certain truth-condition is to be caused by it. But if a belief is caused by one of its truth-conditions, then its truth-condition must obtain, so it must be true.

The mere existence of a truth-condition for an utterance is, in many cases, not sufficient for the utterance to be true. The cause of my belief that, "Lo, a rabbit!" is, presumably, there being a rabbit. But if I mistake a raccoon for a rabbit, then the existence of a rabbit in the next county over won't make my belief true. The truth-condition must not only exist, but also have the right relation to the belief. The causal relation is certainly sufficient for having the right relation, since it's the *causal* relation to the belief that made the cause into the content in the first place.

Note also that there being an appropriate cause for the belief is sufficient for truth, but not necessary. Even when discussing a specific range of beliefs, I'm leery of trying to set up necessary *and* sufficient conditions for something's being true. Any such condition would almost certainly not be as advertised, because, as I tried to show in 1.3, we can't articulate the notion of a truth-condition, and stating necessary and sufficient conditions for something's being true is at least in the vicinity of articulating the notion of a truth-condition.

In the two premises, we can see two themes from chapters one and two. The first premise tries to articulate the idea, present in the theory of interpretation, that cause is meaning. The second premise tries to articulate the idea, present in the theory of meaning, that meaning is truth-condition. By tying causal relations through meaning to truth, we can discover that our beliefs must be true under the right causal circumstances, and that those circumstances are, by necessity, almost always met. Hence most of our observation beliefs are true: only abnormal observation beliefs, brought about by deviant and statistically unusual causal chains, can be false.

The two premises imply:

(3) all of my normal observational beliefs are true.

(1) tells us that a normal observation belief is caused by a satisfier of the truth-conditional predicate of its belief set. But (2) tells us that being so caused is sufficient for being true. So all normal observation beliefs are true. Abnormality is deviance of causal chain, and most observation beliefs with deviant causal chains are false.[54]

To make this argument seem more plausible, I'll now use it against a standard brain-in-a-vat sceptical scenario. As the sceptic would describe it, in the scenario, I am not what, or where, I seem to be. Instead of the usual story I tell about myself, in which my sense organs are sensitive to energy sent from objects that are more or less as science describes them, in the scenario sense organs play no particular role in experience. For

[54] Note that Gettier cases are cases in which a deviant causal chain nevertheless gives rise to a true belief.

sake of the example not being too difficult, though, brains do: the parts of the brain in which my experiences occur are stimulated, not by nervous signals flowing from the sense organs, but rather by various electrodes and other devices connecting my brain to a computer, which stimulates me in accordance with a program designed to give me sensory experiences exactly like the ones I actually have. We may even imagine that the physical laws that the computer and my brain observe are nothing at all like the physical laws that I learned in school, and that 'computer' and 'brain' are only the closest analogies in English to the indescribable reality.

Why is this scenario impossible? — for only its impossibility would allow us to rule it out and defeat the sceptic. Consider the contents of my beliefs. There is a nomic correlation between the computer operating according to a certain algorithm, and my having a certain belief. So an observer would assign the computer's operating according to that algorithm to my belief, as its content. If I have that belief as a consequence of the computer's operating according to the algorithm, then my belief is true, not false. The fact that I don't know much about the subvenient base of my experiences notwithstanding, those experiences prompt true beliefs.

The fact that the brain-in-a-vat scenario is not actually a sceptical scenario is a consequence of the opacity of content. Content is external to the mind; while knowable, it is not knowable "throughout." Mental contents hide quite a lot from the minds of which they are contents. We can explore them to find out more about them, but that process would come to a close only with omniscience.

To make this point clearer, consider the following "sceptical scenario." I think, but am wrong to think, that I am surrounded by middle-sized dry goods. In fact, I am surrounded by tiny particles in fields of force, which generate my illusory experiences. And the way they generate those experiences involve causal laws so mysterious that the closest we can come to understanding them involves the application of concepts like

"wave" and "particle" as analogies to a reality that can't be grasped. Surely, we can't rule out this possibility, and surely, were it actual, our beliefs would be massively wrong!

Even before that description became commonsensical, it would hardly have counted as a sceptical scenario. It only posits that the subvenient base of experience is not yet known, and that reality has as-yet unanticipated features. That we don't grasp the external contents of our thoughts with full transparency doesn't mean that we don't grasp them at all.

We can push the original scenario harder and learn more. What if the computer intentionally generates a new algorithm every time it wants to stimulate a certain "belief"? That is, what if there were no similarity between how it caused the same state in my brain at different times? If, under these antinomian circumstances, an observer couldn't correlate my utterances (never mind how *they* work) with any kind of stimulus, then she would have no basis on which to assign anything at all as the content of my mental states; that is, she would have no basis on which to claim that I *have* mental states. If I'm not interpreted as having mental states, then I don't.

This puts a point rather starkly and paradoxically. Surely it could be the case that things are as the antinomian sceptic describes them, such that no mental state of mine can be nomically correlated with any external circumstance. And yet I would have internal states matching in *feel* every state that I actually have, including my beliefs and even my belief that I have beliefs. It's metaphysically possible for something to have all of my felt experiences, but none of my beliefs. But having a belief is not feeling a certain way. Having a belief is something that can be ascertained from the outside; how you feel is not. The sceptic will find no comfort in the possibility of the antinomian situation he describes, for that situation is one about which the sceptic is wrong. If I have no beliefs, then, while it's trivial that my beliefs are all false, it's also trivial that they're all true. Surely the sceptic has something else in mind: she means to attribute erroneous beliefs, not no beliefs.

Colin McGinn has a reply centering on the rationalization of action by belief:

> Suppose your brain and that of the person on the stimulation machine are sending out impulses causing your respective legs to move in a running motion (your brains, remember, are physically identical). Suppose also that you (but not he) are seeing a tiger running at you with ferocious intent and you believe that this is what you are seeing; you also desire to keep your life. Then we can say that you are intentionally moving your legs in this way because you believe there's a dangerous animal around and you want to escape from it.... But what can we say of your cerebral twin? We cannot say of him what we said of you because he does not *have* experiences and beliefs and desires with those contents. Nor can we rationalize his action by saying that he believes he is being stimulated by an electrode, etc. — for *that* belief would not rationalize his intentionally moving his legs.... (McGinn 1986, p. 361)

It's not clear how my cerebral twin can send out impulses that cause his legs to move, since he has no legs. But we can perfectly well account for my twin's actions, however described. My twin believes, say, that algorithm-17 is active. When algorithm-17 is active, my twin knows, algorithm-73, which tends to trigger pain centers, is typically activated. Also, when algorithm-17 is active, other agents in the vat, those associated in a certain way with the functioning of the algorithm, tend to cease their agency. So, to avoid both algorithm-73's being activated, and the cessation of agency, the agent fires the nerves that tend to de-activate algorithm-17. We might analogically describe this action as "running away."

Admittedly this account is forced. But if we were to redescribe an actual tiger as a pile of particles, we would take away its intensional ferocity just as much we would by redescribing it as an algorithm. Yet a tiger is a pile of particles, and its ferocity remains. Likewise, were a tiger an algorithm, it would be just as ferocious. McGinn seems to want to make mental imagery into the subjects of ferocity:

> The problem here is a general one: we just don't get a coherent, rational, sensible psychology by following Davidson's policy; but we do if we allow the intrinsic properties of the brain a larger role in determining content — in particular, if we attribute the same *experiences* to you and your cerebral twin. (*ibid.* p. 362)

I'm happy to attribute the same experiences, in the sense of subjective feels, to my cerebral twin and me. But I don't run away from the tiger because of the way the tiger

feels to me. I run away from the tiger because I know that it's a tiger, and it's dangerous. Raw feels alone certainly never strike anyone as dangerous; it's only knowledge of dangers associated with objects experienced in a certain way that makes us fear things. In any event, fearing mental images makes no *more* sense than fearing an algorithm, and it seems to me to make a lot *less* sense.

So far, this discussion has focused on the externalistic aspect of content, to the exclusion of the holistic aspect of content. It has also proceeded as though there were a clear distinction between observation beliefs and theoretical beliefs, which there is not. The second phase of the argument focuses on the holistic aspect of content. The first point that needs making is that the holistic aspect of interpretation (and hence of content) guarantees a level of coherence among the interpreted beliefs:

> ...the interpreter will reject a semantic interpretation of a sentence a speaker holds true if the interpretation makes that sentence an obvious contradiction. The interpreter will look askance at an interpretation that finds contradictory two sentences the speaker holds true. Quite generally he must favor interpretations that make the speaker a subscriber to his own, the interpreter's, standards of consistency and rationality, though of course there are times when inconsistency at some point is the best way to accommodate the data. The point behind this policy should be obvious: propositions are identified by the position they occupy among other propositions. If someone seems to have shifted a proposition too far out of position, the reasons for identifying it as *that* proposition will be list.
>
> ...Thought and belief belong to the realm of rationality. Considerable deviations from rationality are consistency with an underlying rationality; but the more extreme the deviations, the less clear it is how the deviations are to be described, and so the less clear it becomes that the norms of thought obtain. (Davidson 1993, pp. 44-5)

The idea here is that the holistic nature of content guarantees that interpreted beliefs will be fairly coherent. Imagine that they were not. Then it's not plain what sort of evidence we could have for attributing those beliefs. We would not be able to generate Tarskian theorems governing the words in the sentences, because every such theorem would give bad predictions in the form of testable axioms.

But if beliefs must be coherent, and a coherent set of beliefs must include a body of (mainly true) observation beliefs, then it's plain that beliefs must be mainly true, for to cohere with a sizable body of truths guarantees truth. Obviously, this view is a form of coherence theory of justification. The basic objection to coherence theories is that coherence is a strictly internal feature of a set of beliefs, one that a set of beliefs can have quite independent of the world. But truth involves a relation to the world. So coherence does not guarantee truth, since it fails to guarantee the right kind of contact between the coherent things and that to which they must correspond to be true. But the present coherentist approach doesn't face that problem. As I tried to show in chapter two, it isn't possible for all beliefs to be theoretical; that is, many of one's beliefs must be observation beliefs, more or less directly triggered by the worldly circumstances that are their meanings. We would have no path to interpretation of a believer none of whose beliefs correlated with immediate external circumstance, so no such believer could exist.[55]

For the purposes of the first phase of the anti-sceptical argument, I needed some characterization of normality of observation beliefs to give sense to the commonsense conditionals in the first premise and the conclusion. For the second phase, I need some further characterization of normalcy for theoretical beliefs. The last paragraph gives me that sense: a normal theoretical belief is one that coheres with the rest of one's beliefs. Most of one's theoretical beliefs have to cohere, else one wouldn't be interpretable and wouldn't have beliefs. So normalcy can be coherence if "normal" takes the sense of "usual." Further, incoherent beliefs are challenges to one's overall interpretation of the person interpreted as having the incoherent beliefs, for incoherent beliefs could, instead, be misunderstood ones. Only coherent beliefs fit into the space of evidence and reasons

[55] see Davidson 1982a for a discussion of the mismatch between what coherence theorists want to justify by coherence — coherent bodies of actual beliefs — and what critics say can't be justified by mere coherence — coherent bodies of independent propositions. The present account maps neatly on Bonjour's Observation Requirement that, for the members of a coherent body of beliefs to be justified, the body of beliefs *must* include beliefs in the reliability of observation beliefs, as well as a whole pile of observation beliefs. (see Bonjour 1985, pp. 140-6)

that are constitutive of the attitudes. So coherent beliefs are also "normal" in a normative sense.

Having this sense of normalcy in place allows me to offer an argument for the truth of all of my normal theoretical beliefs that matches the structure of the first phase of the argument.

(4) For all normal theoretical beliefs of mine,

(a) it is a member of a coherent body of beliefs,

(b) which includes many normal observational beliefs.

The first part is defended with reference to the holism of the attitudes. An incoherent body of beliefs can't exist, and a normal theoretical belief coheres with a body of other beliefs; only statistically and normatively abnormal beliefs are incoherent. The second part is defended with reference to the externalism of the attitudes. Of course, all normal theoretical beliefs of mine also belong to coherent bodies of belief that include no observation beliefs as members; just take some proper subset of my beliefs that includes no observations. But the claim is that there is some set that *does* contain observation beliefs. If a normal theoretical belief were a member of no set of beliefs that contained any observation beliefs, then it would have no worldly contact and hence no content; it would be uninterpretable. (4) can be supplemented by the very obvious:

(5) A belief that coheres with a body of beliefs, many members of which are true, is true.

Coherence is an implicational relation,[56] and implications carry truth. You can't be implied by a group of beliefs that are true, without being true.

Premise (5), or the use to which I put it, is deeper than it looks. The discussion so far appears to indulge in the empiricist conceit that one begins with observations and then builds, foundationalist-style, one's theories on one's observations by way of following out their implications. That is not correct. The coherence that I'm talking about isn't a simple

[56] On coherence, see BonJour 1985, pp. 93-101.

one-way implication of theory by observation. On the contrary, as I tried to show in chapter two, there are no observation beliefs without supporting theoretical beliefs. The coherence of the body of beliefs is that body's having a rich, dense set of implications and other supporting relations, including inductive, abductive, and aesthetic relations. These relations are multi-dimensional, not the one-way relations of foundationalist support.

(3), which tells us that normal observational beliefs are all true, (5), which tells us that to cohere with true beliefs is sufficient for being true, and (4), which tells us that all normal theoretical beliefs do so cohere, jointly imply that all normal theoretical beliefs are true. Only statistically and normatively deviant beliefs are false. Scepticism is mistaken.

How has this argument been an improvement on the omniscient interpreter argument? The omniscient interpreter argument foundered on the possibility of a noniscient interpreter. A noniscient interpreter was universally mistaken, and the noniscient interpreter argument had it that interpretability by the noniscient interpreter was a prerequisite on having beliefs in the first place. But the omniscient interpreter argument, and its dark twin, failed to appeal to the fundamental theses about the determination of content that I discussed in chapter two, externalism and holism. The result of that discussion was that interpreters are partly responsible for someone's having content at all. An interpreter cannot interpret a speaker without grasping a well-confirmed Tarskian truth theory for that speaker, one that correlates her utterances with their causes. If there are no such actual theories held by actual interpreters, then there is no content to be interpreted. To grasp such a theory, the interpreter must be in position to correlate utterance to cause, and hence be correct about the causes themselves. That is inconsistent with being noniscient. So there could be no noniscient interpreter; interpreters, like the interpreted, must be at least mainly right. But interpretation does require massive agreement. So to be interpreted, I must massively agree with my interpreter, who has to be massively right in order to interpret me. The omniscient interpreter argument does

work, but it's unnecessary, as the modal complexities introduced by the notion of the omniscient interpreter add nothing but confusion to the argument, once the deep premises about content have been employed.

Has the argument been a transcendental argument? A Genovan "good" transcendental argument? It has not been a Genovan argument. For Genova, the argument appeals, crucially, to Davidson's opposition to the possibility of alternative conceptual schemes. As can be readily ascertained, the argument that I've presented does not rely on that contention. (To be sure that things could go otherwise than in the order Genova says they go in, I've put off discussion of conceptual relativism until the next section.)

Klein's, then, is the superior understanding of the argument. Klein, though, complained that the argument relied on knowledge of the external world. I don't agree. The fact that appropriate relations between oneself and the rest of the world must obtain for one to be a believer isn't an empirical fact. It's a synthetic *a priori* truth about the attitudes that they are constituted by their causal and inferential relations to worldly events and other beliefs. If I grasp that synthetic *a priori* truth, and also know what my beliefs are, then I know how the world must be for them to exist. That way is such that my beliefs are mainly true.

According to Maker's conception of a transcendental argument, the argument is transcendental. There can be no wild divergence between someone's understanding of the way the world is and the way the world is, not because the mind constitutes its objects, but because the objects constitute the mind as having the contents it does. It's not plain, though, that any substance is left to the notion of a transcendental argument at this point, since any anti-sceptical argument, if successful, would demonstrate the same thing.

Finally, we're now in a position to see the analogy between Kant's and Davidson's refutations of scepticism. For Kant, the mind can identify itself only as against the background of the independent world. But the mind can identify itself: it's possible to know how the world must be for my beliefs to be true. So the independent world must

exist more or less as it's believed to exist. For Davidson, there are beliefs at all only as against the background of causal and nomic relations between the believer and the world, and interpretive relations between the believer and other believers. But, if the sceptic is right, then one has beliefs. So the causal and nomic relations, and the other believers, must exist, and the exact nature of the relationships guarantees that one is generally right in one's beliefs about the world.

3.3 RELATIVISM

Davidson says that, "Conceptual relativism is a heady and exotic doctrine, or would be if we could make good sense of it. The trouble is, as so often in philosophy, it is hard to improve intelligibility while retaining the excitement." (Davidson 1974, p. 183) I concur with this assessment: intelligible doctrines that are called 'conceptual relativism,' 'incommensurability,' and so forth turn out to be commonplaces; interesting doctrines with the same names turn out to be gibberish. Since the idea of conceptual relativism is sort of dated and not, I think, very popular these days, I'll be fairly quick with it.

The notion of conceptual relativity is the notion that the truth of an utterance is relative to the conceptual scheme within which it appears. But, if 'conceptual scheme' just means 'language,' then the point is a truism. Of course the truth of an utterance is relative to its meaning, and it has its meaning only in the language of which it is an utterance. So of course truth is relative to language. I take it that the idea of the conceptual relativists is that conceptual schemes themselves might have a relationship, incommensurability, that means that some truths can't be understood by those who understand other truths. The incommensurability of conceptual schemes, then, seems to be the key to conceptual relativism.

This well-known remark by Kuhn is exciting but of questionable intelligibility: "...we may want to say that after a revolution scientists are responding to a different world." (Kuhn, 1962/1970, p. 111) N. R. Hanson famously offers this little thought experiment: "Let us consider Johannes Kepler: imagine him on a hill watching the dawn.

With him is Tycho Brahe.... *Do Kepler and Tycho see the same thing in the east at dawn?*" (Hanson 1958, p. 5) One of course comes across reference to the conceptual schemes of the Azande and Hopi, so different from ours that we can't understand them. Samuel Delany projects the notion of incommensurable conceptual schemes in science fiction.

In these typical examples, untranslatability is often criterial for conceptual incommensurability. But translation is a terminally linguistic concept. What exactly have conceptual schemes got to do with language? It's a consequence of the publicity constraint on meaning and the triangulation model of content-determination that one has a concept only insofar as one can use that concept in language. Having a concept is essentially a linguistic ability. That lays the groundwork for this remark of Davidson's:

> We may accept the doctrine that associates having a language with having a conceptual scheme. The relation may be supposed to be this: where conceptual schemes differ, so do languages. But speakers of different language may share a conceptual scheme provided there is a way of translating one language into the other.... We may identify conceptual schemes with languages, then, or better, allowing for the possibility that more than one language may express the same scheme, sets of intertranslatable languages. (Davidson 1974, pp. 184-5)

However, translation isn't quite what we want. As Bar-On[57] points out, giving the meaning, in English, of a remark in another language might be a matter of lengthy discussion. Such a discussion isn't really a "translation" as we usually use that word. Better to speak of interpretation. A theorem of a Tarskian truth theory can, in principle, be exorbitantly complicated and yet capture the meaning of the left-hand sentence. Such a theorem is an interpretation. So we should, at a first pass, identify conceptual schemes with inter-interpretable idiolects, and incommensurability of conceptual schemes with impossibility of interpreting one in terms of the other.

But that's not right, either. Bernstein has it that the incommensurability that Kuhn, Feyerabend, and others talk about is what I would call commensurability:

---

[57] See Bar-On, 1994, esp. pp. 150-9.

> ...the "truth" of the incommensurability thesis is not closure but *openness*. For at their best, Kuhn and Feyerabend show us that we can understand the ways in which there are incommensurable paradigms, forms of life, and traditions and that we can understand what is distinctive about them without imposing beliefs, categories, and classifications that are so well entrenched in our own language games that we fail to appreciate their limited perspective. Furthermore, in and through the process of subtle, multiple comparison and contrast, we not only come to understand the alien phenomenon that we are studying but better come to understand ourselves. This openness of understanding and communication goes beyond disputes about the development of the natural sciences; it is fundamental to all understanding. (Bernstein 1983, pp. 91-2)[58]

It's not plain that what Bernstein is describing is all that exciting, but it's certainly intelligible. I think that the most useful way to take Bernstein's interpretation of the idea of incommensurability is that an expression in another idiolect might not be interpretable in my idiolect at a given moment. Incommensurability is a dynamic phenomenon. To commensurate the currently incommensurable, I need to change my own idiolect, possibly by forming new concepts or taking over pieces of the alien idiolect. That is, some idiolects must be changed to become inter-interpretable.[59]

If that's all incommensurability is, then it's intelligible but unexciting. To generate an exciting view, we should treat incommensurability as the notion that there could be rational agents that are terminally and in principle incapable of interpreting one another.[60] One of the agents in principle could never alter his conceptual scheme (other than by abandoning it entirely) in such a way that she could interpret those with the other scheme.

Why "in principle"? As a matter of fact, I won't ever be able to interpret some speakers. I can't interpret Newton, much less Einstein, for I lack calculus; moreover, I can't interpret a calculus textbook. But this sort of contingent failure to learn isn't incommensurability. Incommensurability should be at least slightly exotic.

---

[58] For a similar approach, see Ramberg 1989, pp. 114-37.

[59] Feyerabend (Feyerabend 1987, esp. p. 266) observes that Putnam's critique of conceptual relativism relies on the false assumption that interpretation never requires an alteration of the interpreting language. If this temporal element is really the essence of their notion of incommensurability, then the Gorgiastic cabal gets too excited about it, thinking it profound while it's really just another element of ordinary interpretation.

[60] Perhaps I am "generating" this view, in the sense that no one has ever actually advocated it.

Neil Tennant offers one of the more concrete recent defenses of the idea of incommensurable conceptual schemes. The problem with real-life examples is that, in order to convey the content that is, allegedly, hermetically sealed off from we westerners (Einsteinians, whatever), the defender of incommensurability must state that content in terms that we can understand, thus proving by example that it's commensurable with our conceptual scheme. Tennant turns to hypothetical aliens in the science fiction sense; the ones the Search for Extraterrestrial Intelligence searches for.

Tennant distinguishes between optimistic and pessimistic aliens. Optimistic aliens can be interpreted, though many of our observational beliefs are theoretical for them and *vice versa*.[61] Since incommensurability as I've defined it is uninterpretability, not unusual interpretability, this possibility is uninteresting if we're trying to convince ourselves that there can be incommensurable schemes. The pessimistic possibility is more interesting:

> We might have no way of rendering, in our terms, what would be common to all cases of the same-qualia-for-them. Thus, even prescinding from the problem of trying to communicate what it would be like for them to enjoy such qualia, we might be unable in principle even to specify objective truth-conditions, within our own scheme, for worldly situations that impinge on them in some constant-way-for-them. Thus we might not be able to provide truth-conditions, even using our theoretical conceptual resources, for their observation statements. (Tennant 1999, p. 77)

Undercutting the interest is the phrasing in terms of qualia. Qualia, as terminally private, have no role to play in meaning. The interesting pessimistic possibility is that I might be incapable of identifying truth-conditions for the utterances of an alien, though that alien is speaking.

Davidson's reply to this sort of possibility is not convincing. He says that

> ...nothing, it may be said, could count as evidence that some form of activity could not be interpreted in our language that was not at the same time evidence that that form of activity was not speech behaviour. If this were right, we probably ought to hold that a form of activity that cannot be interpreted as language *in our language* is not speech behaviour. (Davidson 1974, pp. 185-6; emphasis added)

---

[61] See Tennant 1999, pp. 76-7.

The general idea behind this claim is that we have no way to apply the concept of truth except in Tarskian truth theories, which are theories of truth in languages. So we can't apply the concept of truth to a language that we can't interpret with a Tarskian theory, which theory would be in our language. That's true. But the fact that *we* can't interpret a speaker doesn't imply that *nothing* could. The pessimistic aliens' conspecifics can triangulate with them and interpret their utterances.

Narrowly, the pessimistic aliens aren't really a possibility. You don't count as a believer unless you engage in triangulation and interpretation. But to do that, you have to have certain concepts: cause, belief, truth, desire, truth, meaning, and so forth. If the pessimistic aliens really have no beliefs in common with us, then they don't have any of our concepts, including these crucial intensional concepts, and so they don't have beliefs.[62] Could they have only our intensional concepts, but otherwise entirely lack our cognitive apparatus? As a consequence of the holism of the mental, plainly not.

Broadly, it's not plain that this sense of incommensurability is what anyone has in mind. The notion of a scheme is the notion of something that schematizes a content. If two schemes each schematize different content, then it's not plain that they're schemes in the same sense. Two theories of meteorology might differently schematize weather-related phenomena, but there's no sense in which they are alternative schemes to some third, geological, theory. But the pessimistic aliens, by having no belief that we have, have entirely different content; hence their "scheme" isn't really a scheme at all but something different.[63]

Bernard Williams offers a well-regarded defense of something called 'relativism.' For Williams, relativism is possible only when there are at least two systems of belief, each in some sense self-contained, exclusive one of the other, and yet having loci of contact at which the two systems give different answers to the same questions. That there

---

[62] Lepore and Ludwig make this nice, if obvious, point at Lepore and Ludwig, 2005, p. 319.

[63] These last two paragraphs have just drawn out a consequence of the holistic nature of content, which itself is required to respect the intensionality of content attributions.

are these loci of contact shows that incommensurability is not the focus of Williams's relativism. Another feature that the systems must possess is that the second system's giving an answer that is, in some sense, true, even by the lights of the first system, does not give adherents of the first system reason to switch systems. (Williams 1974, pp. 132-5) What does this mean? For scientific relativism, the idea would be that observation underdetermines theory: I can stick with my theory despite your theory's predictions being borne out. For ethical relativism, the idea would be that *behavior* underdetermines theory. I think that Williams is just telling us that "is" does not imply "ought":

> ...the practical question *gets answered* in actual fact, and this occurrence of course trivially satisfies the conditions [on there being a situation allowing for relativism]: the fact that a given question gets answered in this sense in a way that conflicts with, say, the consequences of *S*[ystem]1 does not constrain a holder of *S*1 to abandon his position (he may say that the agent was wrong to so decide). (Williams 1974, p. 136)

This is, of course, not a very interesting idea. The fact that people are sometimes immoral, and sometimes don't know it, has no implications for relativism.

What's much more interesting is what Williams calls[64] the relativism of distance. Williams puts out this form of relativism with reference to two kinds of confrontation between the sorts of systems of belief that can allow for relativism, notional and real confrontations. (Williams 1985, p. 138) We can have a real confrontation with a system of beliefs that we could, in principle, adopt by choice. A confrontation with a system of beliefs that we couldn't really adopt (otherwise than by force, brainwashing, or something else that undercuts the sense of 'adopt') is merely notional. Your typical American, for instance, faces a real confrontation with the socialism-lite of Western Europe. One can envision (or read about) a severe recession and a charismatic leftist changing the structure of our economic system. Your typical American does not face a real confrontation with jihadism. One cannot envision (even in science fiction, if it's to have the least plausibility) any circumstances that would lead Americans, en masse, to adopt violent

[64] at Williams 1985, p. 162

Islamist beliefs. Nevertheless, we can *understand* those beliefs, and, in principle, construct arguments about them; we can even sympathize with them, though we cannot adopt them.

The relativism of distance is this: "Relativism, with regard to a given type of *S*, is the view that for one whose *S* stands in purely notional confrontation with such an *S*, questions of appraisal do not genuinely arise." (*ibid*, p. 142; see also Williams 1985, p. 161) The sense of this relativism is not obvious. Note my example: we can argue about, and even sympathize with, jihadism. But we also think that it's false, and we won't (can't) change our minds. So in what sense can't we assess jihadism? Williams believes that, in fact, our confrontation with jihadism is real:

> Relativism over spatial distance is of no interest or application in the modern world. Today all confrontations between cultures must be real confrontations, and the existence of exotic traditional societies presents quite different, and difficult, issues of whether the rest of the world can or should use power to preserve them, like endangered species... (Williams 1985, p. 163)

This is a strange passage. When faced by an exotic culture, we Westerners do face a real option of destroying the culture, or tolerating it. But believing that traditional societies should be tolerated, or even protected, does not imply that we could, even in principle, accept the beliefs of the culture. So our confrontation does seem to be notional by the announced account of notion-hood; only by this new kind of confrontation, one that asks for tolerance, is the confrontation not notional. A distinction between confrontations in which we are tolerant, and those in which we are not, is orthogonal to a distinction between confrontations in which we might switch sides and those in which we can't.

But allow that notional confrontations are those in which we will never meet a representative of the alternative system of beliefs. We face notional confrontations, then, only with historical or hypothetical systems of beliefs, not with any contemporary systems. Williams's relativism of distance, then, is that it makes no sense to assess systems of beliefs that we can only read about from the past or from speculation.

In a sense, this claim is true. There's something strange about aggressively condemning historical evils — ancient ones, at least — or becoming frantically concerned with the survival of a yet-to-be-realized utopia that one could never join. On the other hand, if we remove the adjectives from what's strange, then there's nothing strange at all. It makes sense to really, deeply hate Nazis, since their atrocities took place in living memory and a movement dedicated to their evil still exists. It makes no sense to really, deeply hate Julius Caesar. That would be strange. But it makes perfect sense to think that Caesar should not have crossed the Rubicon. How can this be a senseless assessment? If the relativism of distance is a comment on the unseemliness of certain overpassionate assessments, then it's reasonable; if it's a serious form of relativism that intends to make some systems of belief incommensurable, then it seems wrong.

Lepore and Ludwig give what I think is the best statement of what's supposed to be exciting in the notions of conceptual schemes and incommensurability: "The excitement of the doctrine of conceptual relativism lies in the thought that, by shifting from the point of view of one conceptual scheme to another, a true sentence may become a false one, while retaining its meaning." (Lepore and Ludwig 2005, p. 318) This sort of conceptual relativism is intended to be intra-linguistic, and it derives, not from anthropological or alien examples, but from examples drawn from the history of science. The idea is supposed to be that, while 'mass' means the same before and after Einstein, still, sentences that include it might change truth value across the revolution. It should be obvious that this sense of incommensurability is unintelligible on my view. Since the meaning of a word is its contribution to the truth-conditions of the sentences in which it appears, same meaning means same truth-conditions. I conclude that there is no exciting, intelligible sense of conceptual relativism. All speakers can, in principle, understand one another.

4 Moral Expressivism

4.0 INTRODUCTION

Moral Realism consists of two claims: that some of our moral utterances are *truth-apt*, and that some of our truth-apt moral utterances are *true*. The usual name for the former thesis is *cognitivism*. The error theorist or moral sceptic denies the second claim, and the task of the balance of this work is to defeat the sceptic and establish the second claim. However, other anti-realist traditions deny cognitivism, and in this chapter I discuss one such tradition, expressivism, primarily as it shows up in the work of a contemporary and important expressivist, Simon Blackburn.[65]

Section 4.1 attempts to state the expressivist thesis clearly. Explaining expressivism isn't easy, because most of the natural ways of stating the expressivist thesis display at least one of two difficulties: 1) they are negative, stating only that assertivism is false, but without actually stating what the expressivist has in mind, or 2) they're patently inconsistent with the main defenses of expressivism against the criticisms to be discussed below. Explaining expressivism is also made more difficult by a lack of clear statements of the thesis by its proponents.

---

[65] I should say something about why I don't discuss other, at least equally important, non-cognitivists, such as Ayer, Stevenson, Hare, and Gibbard. Ayer's discussion (in the first, 1936, edition of Ayer 1946, esp. pp. 102-13) was seminal for spurring discussion of emotivism, which took on a more complete form in the work of Stevenson (see Stevenson 1944 and Stevenson 1963). For precision, though, we should note the deeper historical root in Ogden and Richards 1923/1956, (pp. 124-5, pp. 149-50) perhaps the true source of emotivism this side of Hume. While these historical theorists still maintain some interest (see Wilks 2002), my intent is not chiefly historical and there's precious little of even remotest plausibility in the older theorists that isn't maintained, in a more sophisticated context, in contemporary expressivism. My excuse for ignoring Hare's prescriptivism (on which see Hare 1952) is twofold: on the one hand, Hare's opening moves are so different from those of other non-cognitivists that they would require lengthy discussion on their own but, on the other hand and paradoxically, the objections to expressivism would apply almost unmodifed to Hare's view. Thus the discussion-to-insight ratio would be disadvantageous despite appearances. Finally, no dismissive footnote would be complete without explaining why I avoid Gibbard's well-received and influential theory (on which see Gibbard 1990). Here I admit that the excuses are wearing thin. However, as I think one can see by consulting a symposium on Gibbard (Gibbard 1992a, b, Blackburn 1992, Carson 1992, Hill 1992, and Railton 1992) and other discussions, such as Horwich 1993 and Wedgewood 1997, and comparing the arguments appearing there with my own comments on Blackburn, once again the issues are very similar.

Section 4.2 discusses and evaluates various arguments for expressivism. The main argument for expressivism has to do with the connection between moral utterances and moral motivation. Blackburn tightens this relation beyond what's supported by the facts. I go further and extend the attitudinal holism put forward in section 2.3 to give an alternate account of the facts about moral language and motivation for which the expressivist is trying to give an account.

4.3 and 4.4 attempt to show that expressivism is false by showing that the theory cannot cope with a number of facts about moral discourse and reasoning. I characterize the problems for expressivism as two *embedding problems*. The first embedding problem flows from the fact that moral utterances can be the arguments of truth-functions, such as the material conditional; I discuss this problem in 4.3. Blackburn tries several times to solve this problem by using the *first quasi-realist maneuver*, a reinterpretation of complex utterances within which moral utterances appear so that moral utterances do not appear as the arguments of logical connectives. There are several versions of the first maneuver corresponding to the phases of Blackburn's career, but I'll focus on an *early* first maneuver, in which moral utterances that are seemingly the arguments of logical connectives are taken to refer to the attitudes they would express outside of the seemingly logical context, and a *late* first maneuver involving a modalized reinterpretation of moral discourse. Neither version of the maneuver is satisfying.

The second embedding problem, discussed in 4.4, flows from the fact that moral utterances can appear embedded in contexts like "... is true" or "... is a belief of mine." Blackburn tries to solve the problem by the *second quasi-realist maneuver*, the application of minimalism about truth and belief to moral discourse. This maneuver is substantially deeper than the first maneuver, as, were it successful, it would solve both embedding problems. However, I show that it undermines the sharp cognition/motivation distinction that was to demonstrate expressivism in the first place.

Some opponents of expressivism, such as MacIntyre, might suggest that there is a third embedding problem: moral utterances can also be embedded within rational arguments and they motivate us in ways that they would not if they were not assertions about objective moral facts but rather expressions of speakers' behavioral tendencies.[66] I suspect that there is such an embedding problem. However, this issue is substantially more difficult than the two embedding problems I discuss in 4.3 and 4.4, and it's not clear to me that the expressivist can't make responses that are plausible enough to require extremely lengthy discussion. So I ignore the third embedding problem, though it may be the most important of the three.

4.1 WHAT IS EXPRESSIVISM?

To characterize expressivism, I must begin by characterizing its replacement for cognitivism. I'll use the word 'claim' as a piece of jargon.[67] 'Claim' ranges vaguely but happily over every speech act and mental state — any narrower meaning would either make the formulations of moral realism question-begging or would make redundant the truth-aptness of the moral. For instance, if I were to characterize cognitivism with reference to the truth-aptness of moral *assertions* or *beliefs*, then I would make cognitivism redundant. Assertions and beliefs are truth-apt by nature. 'Claim,' however, through lack of systematic use, may be allowed to vary over cognitive and non-cognitive mental states, and truth-apt as well as non-truth-apt utterances.

I'll begin by characterizing expressivism with reference to language. Though I begin by talking about speech, I move through speech acts to the mental states that, I will say, the speech acts *present*. I say that a speech act *presents* a mental state just in case the

---

[66] See MacIntyre 1984, esp. pp. 6-35. For some discussion, see the sympathetic and critical treatments, respectively, in Lemos 2000 and Unwin 1990. I take the impression (which Lemos confirms at the outset) that MacIntyre's critique of emotivism has not been a focal point of discussion, however often it's mentioned in passing. An interesting difference between Hare's non-cognitivism and expressivism is that Hare's view might have an easier time responding to a MacIntyrean critique. Since Hare's view is a form of Kantian constructivism, it places requirements on moral prescriptions much stricter than those placed on moral expressives by expressivists; the greater strictness might give Hare the argumentative assets required to deal with MacIntyre.

[67] It appears in the canonical statement of moral realism in Sayre-McCord 1988b, 5.

speaker, in performing the speech act, illocutionarily commits herself to the expression of that mental state; or, what I take to be a variant way of saying the same thing, the speech act illocutionarily entails a speech act that would be an expression of the mental state; or, to put it yet a third way, *the mental state is the sincerity condition of the speech act.* [68]

Cognitivism has it that some moral utterances are truth-apt, while any specific non-cognitivism must say just what it is that moral utterances are up to other than being truth-apt. Cognitivism and the alternatives are best characterized, in linguistic terms, with reference to Searle's taxonomy of illocutionary acts.[69] For Searle (as for Austin and other speech act theorists), a speech act is divided into an illocutionary force marker and a propositional content. The initial insight that such a distinction is necessary was, of course, Frege's.[70]

The linguistic focus in earlier parts of this work has been on meanings. But expressivism requires that we consider an additional aspect of linguistic usage: illocutionary force. Davidson himself is no fan of the Searlian approach to illocutionary force that I adopt in this section. I've complained about many things that Davidson has said in previous sections of this work, but on every other issue, I've thought that Davidson was at least in the neighborhood of the right answer, or had gotten lost for plausible reasons. But his account of grammatical mood is hopelessly wrong. I'll discuss my approach briefly while beginning to define expressivism, and then reply to Davidson's objection.

The kinds of speech act with which cognitivism and expressivism are concerned are assertive and expressive speech acts. Here are two sample assertives:

SA1) I have drawn a Queen.

---

[68] On illocutionary commitment and entailment, see Searle and Vanderveeken 1985, 129-60. On sincerity conditions, see Searle 1979, pp. 4-5, and Searle, 1969, ch. 3. but esp. p. 60.

[69] Searle 1979, pp. 1-29. It may seem strange that I state these theses in the vocabulary of speech act theory — especially since this vocabulary seems to be totally foreign to expressivists — but I don't know of any clearer vocabulary to use. An earlier application of speech act theory to expressivism appears in Urmson, 1968, esp. ch. 11.

[70] see Frege 1879, pp. 52-3.

SA2) I wish I had drawn a Queen.

Here are two sample expressives:

SE1) Would that I had drawn a Queen!

SE2) Damn! (Uttered immediately after drawing something other than a Queen.)

The content of SA1 and SE1 are each that I have drawn a Queen. The content of SA2 is that I *wish* I had drawn a Queen. The propositional content of SE2 is that I have *not* drawn a Queen.

The attitudes presented by the assertions are beliefs[71] with propositional contents — meanings, truth-conditions — identical to those of the acts. What makes SA1 and SA2 assertives is that they represent their propositional contents as true.[72] Likewise, what makes the attitudes they present beliefs is that they represent their propositional contents as true. Assertives and beliefs are thus tied together in two ways: each represents its content as true, and, for this reason, assertives are used to present beliefs, rather than some other mental attitude.

What makes SE1 and SE2 expressive is that their point is to express the mental state they present. They make no claims *about* the world (or *on* any person in it). SE1 expresses a non-cognitive attitude (a wish, in this case) toward its content. Likewise, SE2 expresses non-cognitive frustration toward its content. These attitudes are non-cognitive because they do not represent their contents as true; they aren't cognitions (even failed ones) about the world. Expressives are tied to non-cognitive attitudes: neither an expressive nor a non-cognitive attitude[73] presents its content as true. For this reason, the presentation of a non-cognitive attitude is an expressive speech act.

Two small digressions for clarity. Not every expressive speech act presents a non-cognitive attitude. Repeating a catechism, for instance, does not assert the content of the

---

[71] I call all cognitive mental states "beliefs."

[72] Not to be confused with asserting that the content is true.

[73] ...hereafter to be called "desires," with no implicit "mere" or "only" to reduce non-cognitive states to some dubious or questionable state of philosophic disrespectability.

catechism but expresses the speaker's belief in it. Second, though the desire expressed by an expressive speech act obviously *does* make a demand on the world (that's what desires do for a living), the expression of it does not itself make a demand in the same way that some other speech acts, such as orders and promises, do. The illocutionary point of an expressive is to express, not, as with orders and promises, to obligate.

Davidson has no interest in this sort of account of illocutionary force or grammatical mood. On the contrary, he gives an account somewhat similar to his paratactic account of "that"-clauses. The order, "Get back to your room" turns out to be two utterances, both truth-apt: "You will get back to your room. That was in the imperative mood."[74] But the overall utterance is not a conjunction, so it is not truth-apt. It's not clear to me what the overall utterance *is*. It's not a conjunction of two assertions, but what is it instead?

Given the obscurity of Davidson's view, I'm happy to set it aside, but I should still attend to his critique of the illocutionary alternative. Sadly, Davidson doesn't discuss the best account of illocutionary force, but he does discuss related approaches. Of Dummett's treatment of Frege's judgment stroke, Davidson says that,

> ...what bothers me is the implied claim that assertion and the indicative mood can be this closely identified. For there are many utterances of indicative sentences that are not assertions, for example, indicative sentences uttered in play, pretence, joke, and fiction; and of course assertions may be made by uttering sentences in other moods." (Davidson 1979, p. 110)

Our inability to line up assertive force with grammatical mood might lead us to think that the moods were not a good analytical idea. Alternately, we might start drawing some distinctions. For instance, we might distinguish between direct and indirect speech acts.[75] When we make an assertion by way of uttering a sentence with a different mood, we can still have made an assertion by way of uttering a sentence with a non-assertive force, as long as we acknowledge that an utterance can have multiple forces, direct and indirect.

---

[74] Davidson 1979, pp. 119-21.
[75] Searle 1979, pp. 30-57.

For instance, if I say, "I need the salt," while the direct illocutionary force is assertive, the point of my utterance was plainly to get you to hand me the salt, so the utterance has an indirect directive illocutionary force.

We might also distinguish between illocutionary and perlocutionary acts.[76] A perlocutionary act is a non-linguistic act carried out by performing an illocutionary act. I can perform the perlocutionary act of persuading, by way of performing the illocutionary act of asserting. When Davidson says that a joke or a piece of fiction is never an assertion, he has confused perlocutionary and illocutionary acts. I can joke, entertain, or even lie by way of making an assertion. Being insincere does not disqualify an utterance from being an assertion.

The root of Davidson's dissatisfaction with the illocutionary approach is that it fails to satisfy all three of his requirements on a treatment of mood:

> (1) It must show or preserve the relations between indicatives and corresponding sentences in the other moods...
>
> (2) It must assign an element of meaning to utterances in a given mood that is not present in utterances in other moods....
>
> (3) Finally, the theory should be semantically tractable. If the theory conforms to the standards of a theory of truth, then I would say all is well. And on the other hand, if... a standard theory of truth can be shown to be incapable of explaining mood, then truth theory is inadequate as a general theory of language. (Davidson 1979, pp. 115-6)

The illocutionary approach satisfies (1), because the same propositional content appears in the order to clean up your room, as well as the prediction that you will clean up your room, the observation that you have cleaned up your room, and your report that you have cleaned up your room. The illocutionary approach does not satisfy (3). It does not attempt to reduce force to anything semantic. Whether the illocutionary approach satisfies (2) or not is up for grabs. By positing additional elements of an utterance beyond its content,

[76] Searle 1969, p. 25.

does the illocutionary approach posit additional elements of meaning? Or something else?

The demand that a theory of force should be a semantic theory is quite arbitrary. The (depressing to Davidson) conclusion that if a theory of force is not a semantic theory, then the truth-conditional theory of meaning does not exhaust what there is to say about language is only depressing if you began with the expectation that language was limited in features to lexicon, syntax, and semantics. We're not depressed that syntax is not semantic; why should we worry that force is not semantic? *Sans* arbitrary demands on the content of theories about language, Davidson has no case against the notion of illocutionary force.

I return to the main thread. Cognitivism involves the thesis that some moral utterances are assertives, which present cognitive mental states, beliefs. Expressivism is best understood as involving the thesis that (no moral utterances are assertives but that) some or all moral utterances are expressives, which always present non-cognitive mental states.[77]

However, expressivism has a bit of explaining to do at the outset. Consider the moral utterance, "What you did was wrong." The utterance has all of the markers of an assertive. The most obvious interpretation of the utterance is that it has the assertive illocutionary force and that its content is that what you did was wrong. The expressivist, however, will interpret the utterance as an expressive. How?

We may begin with some early remarks by Blackburn. In this early presentation, Blackburn called his view 'projectivism,' and argued that moral utterances involve the "projection" of our moral attitudes. To clarify the notion of projection, Blackburn explains that a hypothetical moral language that "wears its expressive nature on its sleeve" (Blackburn 1984, p. 193) would need

[77] Expressives do not all present non-cognitive attitudes; it's specifically because moral utterances are *moral* expressives that expressivism would say that they present non-cognitive attitudes. There are expressives, such as the repetition of a catechism, that present cognitive attitudes.

> to become an instrument of serious, reflective, evaluative practice, able to express concern for improvements, clashes, implications, and coherence of attitudes.[78] Now one way of doing this is to become like ordinary English. That is, it would invent a predicate answering to the attitude, and treat commitments as if they were judgements, and then use all of the natural devices for debating truth.... *This is what is meant by 'projecting' attitudes onto the world*. (*ibid.* p. 195.)

To project an attitude is to express it through the use of a speech act that *appears* to have the assertive illocutionary force but in fact has the expressive force. Which attitude is expressed is signaled by which "moral predicate" is used. So while the cognitivist (and most natural) interpretation of the example, "What you did was wrong" would have it that the utterance is an assertive the content of which involves the predicate 'wrong,' for the expressivist, 'wrong' does not serve to predicate. One might call it a pseudo-predicate. It answers to the attitude that it expresses, not to some real property (like, say, wrongness) that it's used to predicate. So one may guess that the content might be that you did what you did, or that you did *that*. The attitude expressed is, say, condemnation. The use of the moral pseudo-predicate 'wrong' signals that condemnation is the attitude expressed by the utterance. So for the expressivist, moral predicates serve as illocutionary force indicators as well as markers for the particular non-cognitive attitudes expressed.

I think the foregoing has positioned us to understand expressivism. Expressivism consists in two theses. The first is that the mental states presented by moral utterances are *non-cognitive*. That is the central claim that must be established for us to accept expressivism, and arguments for it will occupy us in the next section.[79] Here, I want to spell out just the one implication of it that's important enough to designate as the second defining thesis of the expressivist. The second claim, expressivism in a narrow sense, will be the linguistic correlate of the first; it is that the illocutionary point of moral utterances

---

[78] Blackburn fails to explain why it is that moral discourse that is overtly expressivist could not serve these functions, and that taking on the guise of truth-aptitude is necessary for moral discourse to serve its functions. Why is it that moral discourse can't look like its true expressive self and still function? Why must it wear the mask of assertiveness and cognitive value?

[79] A prescriptivist who believes that moral utterances are directive speech acts — orders, more or less — will accept the first thesis of expressivism, but deny the second. Most expressivists believe that moral utterances are expressive, but also perhaps directive and sometimes even assertive, though no expressivist can believe that moral utterances are assertive *about their moral content*. I'll leave these complexities aside.

is to express the mental states the utterances present. There's a gap between these claims. It could be that, though we have no moral beliefs but only moralistic desires, nevertheless, in our moral discourse, we're really trying to assert facts. Even though we have no moral beliefs, we talk as though we did. In that case, expressivism about language would be false, though non-cognitivism about moral attitudes would still be true, because, while moral utterances could only present desires, not beliefs, they would still be trying to present beliefs (because the assertion of a fact is the presentation of a belief). So it remains to be seen why the expressivist adopts expressivism (in the narrow sense) in addition to non-cognitivism. The expressivist adopts expressivism because it is necessary to ground his belief that some of our moral speech acts are *fully* successful.

That a speech act is *fully* successful is a matter, I say, of three factors.[80] Let's call the first factor the *performative* factor. Sometimes, I may fail to even perform the speech act I intended to perform. I may try to assert that it is hot in here, for instance, but say, "It's cold in here." In that case, I failed to do what I tried to do. Likewise, however hard I try to perform a marriage ceremony joining an ice cube to a calculator in civil matrimony, my "I now pronounce you man and wife" fails to pronounce anything.

The second factor is the *sincerity* factor. I may, for instance, succeed in performing the act of asserting that *p*, without believing that *p*. In that case, I've lied about *p*, and my act was not *fully* successful. I succeeded in asserting that *p*, but something has nevertheless gone wrong. Likewise, if I promise to *a*, without intending to *a*, then, though I have in fact performed the promise, I was insincere and again was not *fully* successful.

The third factor is the *satisfaction* factor. Assertives are satisfied just in case they are *true*. Orders and promises are satisfied just in case they are *carried*. Not all kinds of speech act have satisfaction conditions; expressives, notably, do not.

---

[80] At least the first factor is of course a massively complex bundle of sub-factors, and all three raise plenty of difficulties that make no difference here.

Consider two conflicting anti-realist directions. Anti-realism will have it that there are no moral facts. Though non-cognitivism has saved our *minds* from error (we have no false moral *beliefs*, for we have no moral beliefs to be false), an anti-realist might say that our *speech* is in error. For when we perform a moral speech act, we perform an assertive act. Such an act fully succeeds only if it is satisfied; it is satisfied only if it is true; but, since there are no moral facts, every such act is not true but false. Hence, none of our moral speech acts fully succeed.

As a rule, there's something wrong, weird, or confused about performing speech acts that, it's known in advance, can't fully succeed. So the anti-realist might worry that the first thesis has undone our moral discourse. That might not be acceptable; we might *need* our moral discourse: it might be important that some of our moral utterances be fully successful. So the expressivist turns another direction. She responds to the worry by denying that when we perform a moral speech act, we perform an assertive act. By turning at this point, the non-cognitivist adopts expressivism. That saves moral discourse by loosening requirements on the full success of moral speech acts. If moral speech acts are expressives, then they are fully successful just in case they are performed and sincere. There's no question that there are moral desires for these expressives to express (that's guaranteed by non-cognitivism), and obviously we perform them all the time (even if, on reflection, we tend to get confused and think that we're performing assertives). So, for the expressivist, our moral speech acts can be *fully* successful even though there are no moral beliefs or facts. The expressivist thus leverages expressivism into the full success of some moral utterances and saves moral discourse from looking erroneous.

4.2 Arguments for Expressivism

In an early work, Blackburn presents three arguments for expressivism that recur through his later work. These are the economic, the metaphysical, and the motivational arguments, and they are the focus of sections 4.2.1, 4.2.2, and 4.2.3, respectively.

4.2.1 The Economic Argument

Considerations of economy yield an *ordering* of meta-ethical theories. According to the expressivist, expressivism is more economical than realism. If we assume that the more economical a theory, the more likely it is to be true, then the expressivist may infer that expressivism is more likely to be true than realism. The argument is thus inductive, in that it only provides *evidence* for some theories and against others. No meta-ethical theory is a deductive conclusion of any economic argument.

Because it's only inductive, the premises and validity of this argument can be accepted at the same time that expressivism is rejected. It may be that, though expressivism is more economical than realism, nevertheless, realism has some other, more important advantages over expressivism. Further, it may be that, though expressivism is more economical than realism, some other theory is more economical yet.

What are the economic advantages of expressivism? Though the argument recurs through Blackburn's later writings, the following early statement is pretty complete:

> [Expressivism] asks no more than this: a natural world, and patterns of reaction to it. By contrast a theory assimilating moral understanding to perception demands more of the world. Perception is a causal process: we perceive those features of things which are responsible for our experiences. It is uneconomical to postulate both a feature of things (the values they have) *and* a mechanism (intuition) by which we are happily aware of it. (Blackburn 1984, p. 182)

We should first consider what ordering is generated. Consider not two theories, but three: cognitivism, scepticism, and expressivism. Cognitivism and expressivism agree that our moral discourse is essentially in order. Cognitivism and scepticism agree that moral claims are beliefs and assertions. Scepticism and expressivism agree that there are no moral facts. Note that cognitivism and scepticism will both "assimilate moral understanding to perception." The cognitivist will think of moral understanding like a working sense like sight, while the sceptic will think of moral understanding like a non-working sense like telepathy. But they will agree that moral understanding is cognitive in structure and function, disagreeing only over whether it succeeds. Scepticism, for this

reason, posits no more than what expressivism posits. It will interpret some mental states as errors instead of non-cognitive attitudes, but it posits no more attitudes or (working) means of cognition than expressivism. So the economic argument begins only by ranking expressivism and scepticism above cognitivism, which is not obviously helpful to expressivism, though it is obviously bad for realism.

The problem here is that the economic argument will work equally well for any non-realist theory, because all non-realist theories refrain from positing the supposedly uneconomical moral properties and cognitions. So the economic argument doesn't really provide even inductive support for expressivism. If successful, it only provides evidence that any anti-realist theory is better than any realist theory, as far as economy goes.

It's not really plain that realism objectionably posits much more than expressivism. Here are the posits, according to Blackburn, of the two theories, with the arrows representing some sort of explanatory relation:

> *Cognitivism*:[81] Shapeless underlying class → shapely property M → perception of it by those with proper affective dispositions ↔ perception of a reason for action → action
>
> *Non-Cognitivism*: Shapeless underlying class → attitudes in those with specific affective dispositions → action (Blackburn 1998, pp. 98-99.)

The first crucial concept here is the "shapelessness" of the underlying class and the realist posit of a "shapely" normative class supervening on the underlying class. Elsewhere, Blackburn explains:

> Let us suppose... that some group of human beings does share a genuine tendency to some reaction in the face of some perceived properties or kinds of things. Surely it need not surprise us *at all* that they should know of no description of what unifies the class of objects eliciting the reaction, except of course the fact that it does so.... Any description is likely to have a partial and disjunctive air.... This may not be a merely practical matter: there is no *a priori* reason to expect there to really *be* a unifying feature. Let us describe this by saying that the grouping of things which is made by projecting our reactive tendency onto the world is *shapeless* with respect to other features. (Blackburn 1981, p. 167)

[81] He has in mind realisms like those of McDowell and Wiggins. See, for instance, McDowell 1981 and Blackburn's criticisms thereof in Blackburn 1981.

A class is *shapeless*, then, just in case its members share nothing in common other than triggering a certain non-cognitive response. Now this makes Blackburn's assumption that the realist and non-cognitivist each start with a shapeless class patently question-begging: according to the realist, the class is normatively shapely. That it isn't, say, physicalistically shapely is a consequence of the irreducibility of the normative to the physical, which is a familiar concept from philosophy of mind. So, with Blackburn's explanation in mind, we have to generate a new characterization of shapelessness that is not question-begging.

Recalling Blackburn's naturalism, we might guess that a class is shapeless just in case its members are not physicalistically similar to one another. And it's surely true that, for at least any normative class I can think of, it is not a physical class nor reducible to one in any way otherwise than very partially and disjunctively. This guess may be supported by the fact that Blackburn's point here is apparently that the shapely normative property posited by the cognitivist has no *explanatory* power, and hence doesn't exist. If we take for granted that only physical classes have explanatory power, and then point out that normative classes are not physical, then we have shown that normative classes lack explanatory power. If we decide that, if a class exists, then it has explanatory power, then we will have decided that there are no normative classes.

There are problems with having a problem with physicalistically shapeless classes. Consider a class that Blackburn does posit, "attitudes in those with specific affective dispositions." This class is shapeless along (at least) two dimensions.

First, dropping the guessed physicalistic aspect of shapelessness, it's not even true that all responses to, say, injustice, are similar *attitudinally*. My response to what I would call 'injustice' might range from cynicism to apathy to irritation to anger to revolutionary fervor. It might even range so far as delight, if I gain a sense of identity from agitating against things I call unjust and I delight in the existence of injustices against which to agitate. Nothing holds together our responses to injustice other than that they are

responses to injustice. If there's no injustice to which they are responses, then nothing holds them together. From a hermeneutical point of view, we need injustice if we're going to group responses to it together. Yet Blackburn does think that this shapeless class of attitudes has explanatory power with respect to action; at least that's what he appears to be thinking when he puts this shapeless class on the left side of an explanatory arrow with actions on the other side.

Second, let's forgive Blackburn this point and assume that every time any person calls something unjust, she has exactly the same attitude toward it: apathy, say. Let's even be forgiving enough to pretend that nothing else triggers apathy, so that the expressivist won't be tempted to misidentify, say, apathy-inducing bad movies as unjust. But what is it that makes my apathy a member of the same class as your apathy, or two of my apathies members of the same class? Apathies are mental states, and classes of mental states are not nowadays widely thought of as physical or reducible to physical classes (or at least not to any such classes that could be identified other than partially and disjunctively). In Davidson's view, the irreducibility of the mental to the physical is a consequence of the holism of the mental. If Davidson's approach to mind is even approximately correct, then the classes of mental states that Blackburn posits are shapeless and as likely to be devoid of explanatory power as realism's normative properties.

There are (at least) two types of consideration of economy, token economy and type economy. Two theories that each posit only members of the same type are not radically different in overall economy however different may be the number of tokens they posit — especially in philosophy, which rarely posits or has much concern with particular tokens but rather with types. But if one theory posits, but another does not, entire types, then there is a sharp difference in economy between the two theories.

Since Blackburn is arguing that anti-realist theories have an advantage over realist theories on grounds of economy, he is considering either a token- or a type-economic advantage. If it's token-economy, then an advantage accrues to the theory that posits the

least, regardless of what the posits are. A theory that failed to posit moral attitudes would have an advantage over expressivism, which would have an advantage over realism. However, given that it's the *shapelessness* of moral properties that makes them objectionable, it appears that Blackburn's argument focuses on type-economic consideration: there is a rule against accepting shapeless types. In that case, a theory that failed to posit moral attitudes would have an enormous advantage over both expressivism and realism, and expressivism would have only a very small advantage over realism.

To be sure, the objectionable class was listed as 'shapely property M,' not as 'shapeless property M.' Perhaps Blackburn has the following in mind. Property M is really just the shapeless underlying property. But, by treating the property as though it were shapely, the realist makes the mistake of treating a shapeless property as though it were shapely. Thus Blackburn can object to the introduction of a shapeless property, at the same time that he attributes to the realist the belief in a shapely property. But this would not be an economy objection. Claiming that a property is shapely, rather than shapeless, is not less economical than admitting that it's shapeless.

Blackburn's comparison of the economies of realism and expressivism contained a further difference. Whereas expressivism posits (physicalistically shapeless) classes of "attitudes in those with specific affective dispositions," realism posits (also physicalistically shapeless) "perception of [normative properties] by those with proper affective dispositions ↔ perception of a reason for action." Here, there is straightforwardly no difference in economy. Realism posits nothing expressivism doesn't also posit. The realist interprets as cognitive the attitudes the expressivist interprets as non-cognitive.

It may be that Blackburn's point is that it's less economical to interpret something as a cognition than it is to interpret it as an affective disposition. If so, he hasn't explained why one kind of interpretation is more *economical* than the other. But I think that Blackburn understands realism as though it were trying to be more like expressivism than

it is. He writes, "...in connection with naturalism, the question to ask of [realism] is why nature should have bothered. Having, as it were, forced us into good conative shape, why not sit back? Why should this [affective moral response] be merely the curtain-raiser for a perceptual system?" (Blackburn 1993, 170) Let me broaden this question, and ask it from the point of view of radical interpretation. Why should we *ever* attribute *any* beliefs? All of our evidence for any attitudinal attribution is behavioral; we attribute attitudes on the basis of observed actions, including speech acts. But all actions are, Blackburn would say, caused by desires, never beliefs. That includes acts of assertion. Why do we need beliefs to help account for actions at all, including assertions? Nature, having gotten everything in order when it came to causing action — conation — needn't have bothered with cognition at all.

Obviously, we aren't going to do away with beliefs with this sort of line. Consider assertion. We interpret an utterance as an assertion because it has the relevant syntactic markers, and when we do so, we also assign a belief in its content. I won't say here why I think that we do that, but it's an intrinsic part of our interpretive practice. Whatever argument we should use to support a belief in beliefs in general is an argument that runs counter to Blackburn's attitudinal economy. We may apply it to moral utterances and achieve the same result for moral attitudes.

Blackburn, however, wrote about a curtain-raiser on a "perceptual" system, whereas I treated him as though he were talking about a "cognitive" system. Perhaps I should take the choice of words more seriously: perhaps Blackburn is wondering why it is that nature needs to give us both a inherent intuition, requiring little or no training, that allows us to perceive moral properties, and also the capacity to be motivated by those properties. Maybe Blackburn's objection here has to with *intuitive* cognition, not cognition in general. This interpretation would fit neatly with another remark I quoted: "Perception is a causal process: we perceive those features of things which are responsible for our experiences." But this version of the argument is just as lame as the

last one. When arguing against the last version, I didn't say anything about whether the assertions for which we attribute beliefs are observation sentences or not. Assume that they are. In that case, why don't we do away with the belief (say) that this is red, and replace it with a spontaneous desire to assert that this is red? The object's being red could cause a desire just as easily as a belief.

For Blackburn, it seems that expressivism is more economical than realism for two reasons. First (though second as I've considered them), realism claims that moral attitudes are cognitive. Second (though first as I've considered them), realism accepts the existence of shapeless normative classes. These reasons correspond to the two theses of realism. The first corresponds to the claim that moral claims are truth-apt, because their truth-aptness is explained with reference to their being cognitive. The second corresponds to the claim that some truth-apt moral claims are true, as their being true is explained with reference to the physicalistically shapeless normative classes. Only the first of these specifically involves expressivism. Positing cognitive attitudes is less economical than not doing so, but we have, in general, rejected this consideration of economy: we reject it every time we attribute any belief at all. The second involves expressivism only insofar as expressivism is a form of anti-realism. It suggests that any form of anti-realism enjoys an advantage over any form of realism. However, it either suggests that expressivism is almost is uneconomical as cognitivism (because it posits equally objectionable shapeless classes), or else that expressivism enjoys no more advantage over cognitivism than a view that did entirely without shapeless classes of moral attitudes would enjoy over expressivism (because expressivism posits some shapeless classes, whereas cognitivism posits more). In neither case does expressivism emerge as a winning view. Expressivism enjoys no more than a very small economic advantage over realism.

4.2.2 The Metaphysical Argument

The metaphysical argument for expressivism deals with the supervenience of the moral on the physical. According to Blackburn, moral properties supervene weakly, but

not strongly, on physicalistic properties.[82] Expressivism can explain this fact, realism cannot. Expressivism therefore has greater explanatory power and, if we assume that truth tracks explanatory power, expressivism is more likely to be true than realism. As with the economic argument, the metaphysical argument is inductive. It may be that, though when it comes to supervenience, realism has less explanatory power than expressivism, realism has greater explanatory power elsewhere. It may be that, while expressivism has more explanatory power than realism, nevertheless some other theory has yet more. So we could, in principle, accept the soundness of the metaphysical argument without accepting expressivism. As it turns out, we need not accept the premises.

Here are claims for the supervenience relations in question:

> WEAK: Within a given world, if x and y are identical with respect to their non-moral properties, they are identical with respect to their moral properties.
>
> STRONG: If x and y are identical with respect to their non-moral properties, they are identical with respect to their moral properties, even if x and y are in different worlds from one another and/or the speaker.

According to Blackburn, WEAK is true, but STRONG is not. Expressivism can explain WEAK, while realism cannot.

Any theory that committed itself to strong supervenience would gain weak supervenience as a trivial consequence: STRONG implies WEAK. So it may be that a realism can explain weak supervenience with reference to strong supervenience. But for Blackburn, it's plain that the moral does not strongly supervene on the non-moral:

> It does not seem a matter of conceptual or logical necessity that any given total natural state of a thing gives it some particular moral quality. For to tell which moral quality results from a given natural state means using standards whose correctness cannot be shown by conceptual means alone. It means moralizing, and bad people moralize badly, but need not be confused. (Blackburn 1984, p. 184)

If STRONG is true, then claims of the following sort would be necessarily true (if true):

---

[82] Blackburn does not use the vocabulary of strong and weak supervenience. I'm borrowing it from Kim, 1984.

For anything with non-moral properties $F_1...F_n$, that thing has moral property G. This claim is not known to be true as a matter of conceptual or logical analysis. From this premise, Blackburn then leaps to the conclusion that the claim is not a necessary truth. It seems that Blackburn regards necessity as at least extensionally equivalent to analyticity (being logically true) and *a prioricity* (being conceptually true). But it's now widely accepted that that there are synthetic, *a posteriori*, necessary truths. Unless this currently fashionable view is false, Blackburn's argument is a simple *non sequitur*. So far as I know, he's given us no reason to reject the fashion.

However, it would be unfortunate if realism relied on STRONG: we wish to assume as little as possible. So it's important to see why it is that realism is, without invoking STRONG, still no less able to account for WEAK than expressivism is. A good way to start on this is to look at Blackburn's account of why the expressivist has such an easy time accounting for WEAK. But before we can look at his account, there's a problem. Expressivism, as a version of anti-realism, denies the existence of moral properties. For the expressivist, there are no moral properties to supervene on non-moral properties. How can anyone account for the supervenience of properties that don't exist?[83]

One way would be to introduce moral properties in some allegedly special, non-realistic way, while maintaining that moral properties introduced in this special way don't violate the original anti-realism. I'll discuss such a project in 4.4 and find it wanting. Here, I'll do something less ambitious: I'll come up with some other relevant properties, that are not moral properties, but that are such that the expressivist can explain their weak supervenience on other non-moral properties. To do this explaining, we should appeal directly to expressivism. If a speaker performs a moral utterance, such as "Murder is wrong," she expresses her desire that murder not happen. Such an utterance might prompt us to look at the properties of murders and see whether any of them explain why the

[83] I follow my routine of using 'property' and 'predicate' more or less interchangeably; when I affirm that some properties supervene on others, I mean either that some universal supervene on others, or else that some predicates' application conditions have a supervenience-like relation to other predicates.

speaker has this attitude, and the property of causing this sort of non-cognitive attitude might be what it is that supervenes on other non-moral properties of murder. So expressivism might not commit itself to WEAK, exactly, but rather to:

> QUASI-WEAK: Within a given world, if x and y are identical with respect to their *other* non-moral properties, they are identical with respect to their non-moral properties of causing moralistic desires that they (i.e., x and y) occur or exist.[84]

This is perhaps not a spectacularly lucid statement of the thesis. The idea is that, whereas the realist can divide properties into normative and non-normative ones, and claim that normative properties supervene on non-normative ones, the expressivist can't do the same because the expressivist doesn't accept normative properties. So the expressivist, in order to defend any supervenience thesis, must introduce a distinction among the non-normative properties. The distinction is between the property of causing moral attitudes, and all of the other non-normative properties. The thesis is that if one thing causes a certain moral attitude, and a second thing is like the first in every way *other* than causing moral attitudes, then it is *also* like the first in causing that moral attitude. To allow for bad moralizing and hypocrisy, we can allow QUASI-WEAK to be defeasible or otherwise soft enough to handle contingencies.

Why should QUASI-WEAK be true? Blackburn explains that, "Our purpose in projecting value predicates may demand that we respect supervenience." (*ibid*, p. 186) Why would that be? Because when we perform a moral utterance, "We intend coordination with similar avowals or potential avowals from others, and this is the point[85] of the communication. When this coordination is achieved, an intended direction is given to our joint practical lives and choices." (Blackburn 1998, pp. 68-9) The idea appears to be that, if our moral attitudes were not systematically related to non-moral properties of things (other than their non-moral power to cause moral attitudes), then the point of having moral attitudes would be thwarted. If I don't value consistently, I can't coordinate

---

[84] STRONG implies WEAK. But neither STRONG nor WEAK imply QUASI-WEAK.

[85] Coordination is not the *illocutionary*, but the *perlocutionary*, point. Blackburn's claim is not inconsistent with my formulation of expressivism in the terms of speech act theory.

my actions with those of others, and so I may as well not have the moral attitudes in the first place.[86] Expressivism explains QUASI-WEAK with reference to the purpose of having moral attitudes.

The tasks facing the realist and the expressivist are incommensurable. The realist must explain WEAK, while the expressivist must explain QUASI-WEAK. However, the two theses play similar roles in their respective overall theories, so the burden on the realist doesn't seem sufficiently different from the burden on the expressivist to wreck the metaphysical argument, which is supposed to show that expressivism has *more*, not just *different*, explanatory power from realism.

How might a realist explain WEAK? The trouble with using the metaphysical argument to show that expressivism is superior to any version of moral realism is that there are many versions of moral realism, and some of them may have a better time of explaining WEAK than others. To explain WEAK, one would have to articulate a metaphysics of morals. This isn't the place for that, though I make an approach on this issue in 5.2.3. The metaphysical argument invites a philosophical set-piece in which the realist articulates a complete realistic theory to explain the supervenience. But no such set-piece is necessary to defeat the metaphysical argument. The metaphysical argument is a comparative argument, and to defeat it it's sufficient to show that expressivism has no more explanatory power than realism. One way of doing this is to show that expressivism has no explanatory power. So, instead of showing that realism *can* explain WEAK, I'll show that expressivism *can't* explain QUASI-WEAK — and in the bargain, show that expressivism would probably imply something like STRONG if it did imply QUASI-WEAK. While this extra implication isn't objectionable to everyone, Blackburn would find it objectionable.

[86] I assume that we needn't take seriously Blackburn's exact wording: it's obviously false to say that the point of moral utterances is to coordinate *utterances*; surely the expressivist thinks that the point is to coordinate actions in general.

For Blackburn's expressivism, the account of QUASI-WEAK is that moral utterances aim at coordination. But if we explain QUASI-WEAK with reference to our needs for coordination, then QUASI-WEAK should apply only where we have a need to coordinate. To make clear the difference between what the expressivist wants and what she can have, I'll rephrase QUASI-WEAK (and rename it, BROAD QUASI-WEAK) and introduce a new thesis, NARROW QUASI-WEAK:

> BROAD QUASI-WEAK: Within a given world, if x and y are identical with respect to their *other* non-moral properties, then, even if no coordination is necessary with respect to x and y, they are identical with respect to their non-moral properties of causing moralistic desires that they (i.e., x and y) occur or exist.
>
> NARROW QUASI-WEAK: Within a given world, if x and y are identical with respect to their *other* non-moral properties, then, if coordination is necessary with respect to x and y, then they are identical with respect to their non-moral properties of causing moralistic desires that they (i.e., x and y) occur or exist.[87]

Two small points for clarity. First, it's obvious that whether coordination is necessary with respect to x and y is not to be one of the "other non-moral properties" in either thesis. Second, 'desire' is a generic term for a motivational attitude. On this use of the term, I can desire that something happened, or not have happened, in the past. For instance, I desire that William have been beaten at Hastings. Also, I can desire that something happen or not in a future that I believe won't be realized; for instance, I desire that the Greens beat the Reds in Kim Stanley Robinson's science fiction novels about the settlement of Mars, even though I don't believe that there will be Greens and Reds.

Expressivism can account for NARROW QUASI-WEAK. For expressivism, the purpose of moral attitudes is to coordinate action. Since, in NARROW QUASI-WEAK, coordination with respect to x and y is necessary, and having attitudes that failed to respect NARROW QUASI-WEAK would make the attitudes non-coordinating, having

[87] QUASI-WEAK implies BROAD QUASI-WEAK (through being essentially the same thesis), which implies NARROW QUASI-WEAK. Neither STRONG nor WEAK imply or are implied by any of the QUASI-theses.

attitudes that failed to respect NARROW QUASI-WEAK would defeat the point of having the attitudes. So expressivism can account for NARROW QUASI-WEAK.

However, where coordination is not necessary, neither is supervenience. For instance, you and I do not need to coordinate on the subject of whether Julius Caesar should have crossed the Rubicon, or on what either of us should do in private. However, even though I don't need to coordinate with regard to Caesar, I can have a negative attitude toward his treachery, and that attitude needs to line up with my attitudes about other people's non-morally similar actions. Likewise, though you and I do not need to coordinate on private matters, we often have attitudes about the moral qualities of others' private acts. These attitudes and condemnations don't serve the goal of coordination, so, on expressivism, there's no reason they should respect the supervenience of the moral on the non-moral.

It's true that we rely on facts about the past to guide us in future moral decisions, and that we rely on our judgments about similar private cases in making judgments about our own private behavior. But why is that? We feel comfortable relying on the past and the wider world as a guide just because we accept the supervenience of the moral on the non-moral. So we need to coordinate about the past and matters not of mutual concern because we already accept the supervenience of the moral on the non-moral. Blackburn's approach inverts the direction of explanation. For Blackburn, the need to coordinate explains supervenience. But, in lieu of supervenience, we only need to coordinate our attitudes toward future actions of mutual concern. So, in lieu of supervenience, we would have no coordinative need that would account for the supervenience of the moral on the non-moral in the past, or outside a realm of mutual concern. So there is no expressivist explanation for the supervenience of the moral on the non-moral in the past, or outside of a realm of mutual concern. Yet supervenience holds in those areas. Expressivism thus has no explanatory advantage over realism.

To be sure, Blackburn did say that the point of moral utterances was "coordination with similar *avowals* or potential *avowals* from others," but this claim is obviously false. We don't talk so that we can coordinate our talking, we talk so that we can coordinate everything that isn't talking. If all we could coordinate were our moral utterances, then it would be false that, through coordination, "an intended direction is given to our joint practical lives and choices." Our joint practical lives and choices extend beyond our moral utterances. Further, the point of coordinating attitudes is only to coordinate behavior. Where we have no need to coordinate behavior, we have no need to coordinate attitudes. So I assume that Blackburn's wording was off and that expressivism does not involve the thesis that our moral talk has no extralinguistic purpose.

To add insult to injury, I want to point out that, if the expressivist *could* account for QUASI-WEAK, she might also account for:

> QUASI-STRONG: If x and y are identical with respect to their *other* non-moral properties, they are identical with respect to their non-moral properties of causing moralistic desires in the speaker that they (i.e., x and y) occur or exist, even if x and y are in different worlds (from one another and/or the speaker).[88]

Why? Because the expressivist believes that coordination is the point of moral attitudes, she is in fact limited to explaining supervenience as it applies to moral discourse where coordination is necessary. But, if she could break this limit and account for supervenience where coordination is not necessary, then there's no reason to think that she wouldn't account for supervenience across worlds. We don't need to coordinate on the subject of events that won't take place but could (e.g. as of this writing, Nader's inauguration), just as we don't need to coordinate on events in the past (e.g., Caesar's crossing the Rubicon) or that otherwise don't involve us (e.g., private behavior). The expressivist who rejects STRONG or QUASI-STRONG would need to explain why supervenience applies to some moral discourse about which no coordination is necessary, but not others. We can argue about the merits of a Nader inauguration, knowing that there will never be one, just

[88] QUASI-STRONG implies all of the other QUASI-theses, but it neither implies nor is implied by either STRONG or WEAK.

as we can argue about the merits of Rubicon-crossing, though our attitudes don't help coordinate anything, or a private act, though our attitudes have no effect on whether the act occurs and hence our coordination (and presumably attitudes) is (are) pointless.

The metaphysical argument is a comparative argument that tries to show that expressivism is more likely to be true than realism because it can explain the weak supervenience of the moral on the non-moral. To defeat this argument, it's sufficient to show that expressivism has no more explanatory power than realism, and to show this point, it's sufficient to show that expressivism has little or no explanatory power when it comes to the weak supervenience of the moral on the non-moral. Since the expressivist explanation only applies where coordination is necessary, and yet there is moral discourse on matters about which coordination is not necessary, the expressivist explanation fails. With respect to explanatory power, as with respect to economy, expressivism enjoys no advantage over realism.

4.2.3 The Motivational Argument

The motivational argument for expressivism is also the motivating argument for expressivism; this is the argument which recurs most through Blackburn's later writing, and which has the deepest roots in the history of philosophy.

The motivational argument begins with the standard belief-desire account of action. Given an agent's belief that the glass contains caffeine, and her desire to drink caffeine, we have an explanation for her drinking the contents of the glass. But neither any belief nor any desire could account for any action on its own.[89] So, if moral utterances present beliefs, then those beliefs require supplementation by desires before

[89] This may be too simple. A simple action, engaged in as an end in itself, could be explained by a desire absent any beliefs. If I desire to wiggle my toes, that desire is probably sufficient on its own to explain my wiggling of my toes. The desire might need supplementation by beliefs like that wiggling my toes will not cause other things I desire not occur (e.g., nuclear war), but this sort of supplementation seems, on face, different from the sort of supplementary belief that tells you that this action will help you fulfill your desires.

they can help explain moral behavior. Likewise, if they present desires, then those desires require supplementation by beliefs before they can help explain moral behavior.

The discussion of belief-desire explanations is to help show that moral utterances present desires:

> It seems to be a conceptual truth that to regard something as good is to feel a pull towards promoting or choosing it, or towards wanting other people to feel the pull towards promoting or choosing it. Whereas if moral commitments express [i.e., moral utterances present] beliefs that certain truth-conditions are met, then they could apparently co-exist with any kind of attitude to things meeting the truth-conditions. Someone might be indifferent to things which he regards as good, or actively hostile to them. (Blackburn 1984, p. 188)

The argument is that, if moral utterances did not present desires, then there would be no explanation for our attraction to what we call 'good' or 'right.' The mental states presented by our (sincere) moral utterances line up very neatly — not perfectly, but almost — with our behavior. Usually, for any *x* that a speaker calls 'good' or 'right,' that speaker behaves in such a way as to bring it about that *x* exists or occurs. This would be natural if calling something 'good' or 'right' expressed the desire that it exist or occur. But if calling something 'good' or 'right' only presented the *belief* that it is good or right, then there would be no such explanation. Why can't beliefs incline us toward moral action? Because "...we have no conception of a 'truth condition' or fact of which mere apprehension by itself determines practical issues. For any fact, there is a question of what to do about it. But evaluative discussion just is discussion of what to do about things." (Blackburn 1998, p. 70) Unlike the economic and metaphysical arguments, the motivational argument is a deductive argument with expressivism as its conclusion. If it is sound, then expressivism is true, so I want to lay it out a little more precisely. Here is the motivational argument as I understand it:

> 1) The mental states presented by our moral utterances help to explain our moral behavior.
> 2) The mental states presented by our moral utterances help to explain our moral behavior if and only if they are motivational.
> Thus, C1) The mental states presented by our moral utterances are motivational.
> 3) No motivational mental state is cognitive; they are all non-cognitive.

Thus, E1) The mental states presented by moral utterances are non-cognitive.

The validity is obvious, but of the premises, none is uncontroversial.

How might the realist deny 2? To deny 2, the realist must accept that the mental states presented by our moral utterances do help to explain our moral behavior, but deny that they need to be motivational in order to play their explanatory role. Consider this explanation of an action:

> Des: that I drink water
> Bel: that if I drink the contents of this cup, then I drink water
> Explains (*ceteris paribus*)
> Act: that I drink the contents of this cup

That's pretty straightforward. No one would be tempted to suggest that, were the agent of this action to say, "If I drink the contents of this cup, then I drink water," that the agent had expressed some non-cognitive attitude about the cup and water. But it's not obvious that we should react any differently to this explanation:

> Des: that I do what's most good
> Bel: that if I do *this*, then I do what's most good
> Explains (*ceteris paribus*)
> Act: that I do *this*.

Why is the expressivist tempted to claim that, if this agent says, "If I do this, then I do what's most good," or "This is the most good thing to do," the agent thereby does not assert a fact but rather expresses a non-cognitive attitude? What's the difference between the two cases?

When we deny 2, we can see many ways in which the attitudes presented by moral utterances could help explain action. They could motivate by being beliefs that an action is good, which work in tandem with a desire to do the good. Or each one could cause and rationalize a matching desire. Or each one could cause and rationalize a matching intention; there's no reason our motivationally powerful attitudes need to be desires. Blackburn, however, thinks that the posit of a desire for the good is a concession to the expressivist:

> ...we seem to be faced with the following zig-zag structure. We start with something bad such as the piano being on your foot. An agent is concerned about that. ...But the philosopher worries that this is very *mere* attitude or emotion on the part of the agent. So she substitutes, as the focal point of ethics, that the agent believes that she ought to help — a belief that is true, and carries the authority of truth. So if an agent fails to believe that she ought to help, it would be an Apollonian defect, not merely one of Dionysus. But then, alas, there are immoral and amoral agents, who know this truth and do not care about it. Are they 'merely' wrong on the Dionysian side? Perhaps we had better find a property of the moral belief that should sway them; for instance, that it is reasonable to be swayed by it and only it. ...But suppose they do know that it is unreasonable, but just don't care about that? ...Eventually, we will be bound to finish by appealing to Dionysus for help: 'But *don't* you care about things going well, or flourishing or social coordination, or peace, or contracts?' Fortunately, most agents do care.... All the zig-zags have been a delaying tactic, for, as Hume saw, somewhere there will always have to be a practical, dynamic state. (*ibid*, p. 91)

Perhaps Hume is right, but the fact that there must *somewhere* be a desire that's relevant to the explanation of moral behavior does not imply that that desire must be somewhere *in particular*. It needn't be the one presented by a moral utterance. The zig-zag structure would be a delaying tactic for a realist who wants desires to play *no* role in morality. This sort of realist, who tends to put 'mere' before 'desire,' has an allergy to allowing desires to play a role in the explanation of moral behavior, but a realist doesn't have to have this allergy. The fact that the realist believes that moral utterances are on the side of Apollo doesn't mean that she must claim that morality doesn't in any *other* way involve Dionysus. Furthermore, Blackburn needs to account for his choice of motivating attitude: why desire? Why must the "practical, dynamic" state be one of "caring," with its attendant features extraneous to its motivating function?

However, I've not yet offered an explanation of the motivational power of moral claims. I've shown that it's *possible* that the mental states that incline us to behave ethically line up with our moral utterances. But why should they?

The answer flows from the holistic constraints on radical interpretation:

> ...we could not begin to decode a man's sayings if we could not make out his attitudes towards his sentences, such as holding, wishing, or wanting them to be true. Beginning from these attitudes, we must work out a theory of what he

> means, thus simultaneously giving content to his attitudes, and to his words. In our need to make him make sense, we will try for a theory that finds him consistent, a believer of truths, *and a lover of the good*... (Davidson 1970, p. 222, emphasis added)

In previous chapters, I've discussed holism as it applies to attitude attributions and semantic content attributions. Here, I should narrow the focus and look at holism as it applies to normative utterances and motivational states.

Consider the procedure of radical interpretation. We're proceeding as normal, hypothesizing T-sentences, inventing axioms to account for them, generating new theorems to test the axioms, often revising the axioms and occasionally rejecting the evidence. We suspect that a good axiom for '↗' might be: for any x, x satisfies '↗' iff x is good.

The fact that we're necessarily motivated by our moral convictions suggests that more will be necessary for us to check this axiom than is necessary for some other axioms. For instance, the fact that our informant systematically agrees that good things are ↗ will provide excellent evidence for the axiom, but what if she never does anything that she regards as ↗? We would count this as very strong disconfirming evidence for our axiom. No predicate can mean what 'good' does unless believing that something satisfies that predicate (somehow) inclines the believer toward it. While 'good' predicates are ordinary predicates that appear in cognitive judgments, they are by nature closely connected with motivation. A speaker's attitude toward an act cannot be captured by our concept of the good unless the speaker is inclined toward the act. I provide a more detailed argument along these lines in 5.2.4

Considerations like this lend support to the idea that, to speak a language, one must be a lover of the good (in some, perhaps passionless, sense of 'lover'). Since one must be inclined toward what one believes to be good to count as having such beliefs, the belief that something is good need not itself motivate one toward good behavior. The necessary connection between believing that something is good, and wanting it, is

provided by a holistic constraint on radical interpretation of normative concepts, not by the belief's really being a desire in disguise.

This account of the connection between motivation and moral conviction is superior to the expressivist account. As I'll show in later sections, expressivism has large explanatory obligations, and fails to fulfill them. This account, on the other hand, flows from an attractive attitudinal holism that stems from general constraints on interpretation. My holistic account, then, takes a very large lead on expressivism, at least if the arguments of sections 4.3 and 4.4 are correct.

4.3 THE FREGE-GEACH PROBLEM

The Frege-Geach problem is the problem of embedding moral utterances in complex logical sentences once they have been construed expressivistically, as not truth-apt.[90] For instance, consider the expressivist interpretation of an utterance like, "Your action was wrong." For the expressivist, 'wrong' is a pseudo-predicate the linguistic function of which is to mark the attitude being expressed toward the subject of the sentence. If we represent the cognitivist interpretation of the utterance like this, using the Fregean 'judgment stroke' to indicate assertive illocutionary force:

⊢ that your action was wrong

...then we can represent the expressivist interpretation of the utterance like this, using 'E' to represent the expressive illocutionary force:

E contempt: that you performed your action.

So far, no problem. But we should ask how expressivism manages utterances like these:

Ex1) If your action was wrong, then you should be punished.

Ex2) If your action was murder, then your action was wrong.

Ex3) If your action was wrong, then you will be punished.

---

[90] The seminal presentation of what's come to be known as the Frege-Geach problem is Geach 1965; earlier, related discussions are Geach 1960 and Searle 1962. See Price 1994 for discussion. Price, unfortunately, tries to help the expressivist avail himself of the same minimalist solution that section 4.4 will show to be unavailable to the expressivist.

Each of the three conditionals has at least one moral pseudo-predicate; Ex1 has two ('wrong' but also 'should').

First, how does the cognitivist deal with these contexts? What, for the cognitivist, do the embedded moral propositions mean, when embedded? The cognitivist draws immediately on the Fregean lesson that neither the antecedent nor the consequent of a conditional is asserted when it appears in the conditional. The assertive illocutionary force attaches neither to the antecedent nor to the consequent, but to the conditional as a whole. The propositional content of the assertion is logically complex:

⊢ (your action was wrong → you should be punished)

The cognitivist can offer this interpretation because of the nature of the assertive illocutionary force. The assertive force operates directly over contents. Those contents can be embedded in logical complexes that can be asserted. So the content is the same, whether embedded or not.

But the expressive force doesn't work that way. What's expressed is not a content, but an attitude toward a content. Expressive speech acts are more complicated than assertives, because, while both assertives and expressives represent contents, expressives also express attitudes toward those contents. Assertives do not express attitudes, not even beliefs (though they do present them, since that one believes the content of an assertion is the sincerity condition for the assertion). Likewise, if a moral utterance appears as the antecedent or consequent of a conditional, it is not being used to express an attitude. Expressivism, then, has to explain how a speaker can perform an expressive act without expressing anything.

Attitudes can appear in language in two ways. First, they can be expressed. Second, they can be referred to. Blackburn's first (but also his most recent) attempt to solve the Frege-Geach problem suggests that sometimes, utterances that would ordinarily express attitudes refer to them instead. Such an approach appears by implication in *Spreading the Word* (pp. 193-5), but explicitly in the more recent *Ruling Passions*:

> Suppose I say that the sentence 'Bears hibernate' expresses a belief.[91] Well, it only does so when the sentence is put forward in an assertoric context. So what happens when it is put forward in an indirect context, such as 'If bears hibernate, they wake up hungry'? For here no belief in bears hibernating is expressed. The standard answer is to introduce a proposition or thought, regarded as a constant factor in both the assertoric and the indirect context. When we say bears hibernate, we express or assert the proposition,[92] and represent ourselves as believing it; when we say 'If bears hibernate...' we introduce the proposition in a different way, conditionally, or as a supposition. Frege thought that in this second kind of context we refer to the thought that we assert in the assertoric context. (Blackburn 1998, p. 71)

This leaves a lot to be desired as Frege interpretation. Blackburn's errors begin with his failure to grasp the difference between two contexts in which a proposition appears unasserted: when it appears embedded in a logically complex utterance, and when it appears in an indirect context, such as those defined by "...is true," "I believe that...," and so forth. In the latter, indirect, cases, Frege claims that we refer to the thought that we would have asserted with the same utterance, had it appeared in a direct context. But when I say "If the sky is blue, then grass is green," I have referred to neither the thought that the sky is blue, nor the thought that the grass is green. I have *referred* to no thoughts at all, though I have *asserted* a complex thought that's partly composed of the thoughts that the sky is blue and that grass is green.

Even in the indirect contexts, we don't refer to *attitudes*. If I assert that, "Bob believes that grass is green," I haven't referred to Bob's belief. I've referred to the thought that grass is green, but not to Bob's belief in that thought. I've asserted that Bob has such a belief. Blackburn is trying to Frege his way out of a Frege-Geach, but his proposal is in fact entirely novel:

> If this [the aforementioned pseudo-Fregean story] is allowed to solve the problem for ordinary beliefs, it might simply be taken over by the expressivist. In the Fregean story a 'proposition' or 'thought' is simply introduced as the common element between contexts: something capable of being believed but equally

---

[91] Notice that on this sentence's use of 'express,' the objects of this verb are *attitudes*, such as belief.

[92] Notice that in this sentence, 'assert' and 'express' appear to have the same sense, and the objects of both verbs are *propositions*, not attitudes. It would be nice if expressivists could use the word 'express' in some sort of consistent, clear way.

> capable of being merely supposed or entertained. So why not say the same about an 'attitude'? It can be avowed, or it can be put forward without avowal, as a *topic* for discussion, or as an alternative. Just as we want to know the implications of a proposition or a thought, we want to know the implications of attitude. What implies it, what is it right to hold if it is adopted? (*ibid*, emphasis original)

This explanation leaves a lot to be desired. It's not plain that attitudes have implications; in fact, whether attitudes have implications (or are implied) is sort of the problem to be addressed.[93] It's not plain what it is to put forward an attitude one doesn't have, or, indeed, to put forward an attitude. But, since it doesn't mean to express the attitude, we may guess that it means to refer to the attitude.

We can gain more insight by returning to the original presentation from *Spreading the Word*:

> ...the first thing we need is a view of what we are up to in putting commitments into conditionals. Working out their implications, naturally. But how can attitudes as opposed to beliefs have implications? At this point we must turn again to the projective picture. A moral *sensibility*, on that picture, is defined by a function from *input* of belief to *output* of attitude. Now not all such sensibilities are admirable.... And amongst the features of sensibilities which matter are, of course, not only the actual attitudes which are the output, but the interactions between them. For instance, a sensibility which *pairs* an attitude of disapproval towards telling lies, and an attitude of calm or approval towards getting your little brother to tell lies, would not meet my endorsement. (Blackburn 1984, p. 192)

As usual, we're adrift in seas of uncharted language. Moral utterances are here *commitments*, but there's no word on what a commitment might be. We have to guess 'expressive' and move on. Likewise for *endorsement*, and again we must guess 'expressive.' A 'pairing,' presumably, is just two attitudes had by the same person. Somehow, though a sensibility is a function from beliefs to non-cognitive attitudes, a good example of the workings of sensibility deals with sensibility as a function from non-cognitive attitudes to other non-cognitive attitudes. Presumably, sensibility can take

---

[93] It's worth noting again the difference between Frege's account and Blackburn's: for Frege, we entertain, not beliefs, but thoughts, whereas for Blackburn, we entertain attitudes toward thoughts. On the general issue of expressivism and logical consistency, see Tersman 1995, van Roojen 1996, and Björnsson 2001. van Roojen's discussion included an historical perspective that helped me to see the differences between Blackburn's positions across time. Tersman's paper links the issue of consistency with the issue of supervenience from section 4.2.

either cognitive or non-cognitive attitudes as arguments, and yields non-cognitive attitudes as values.

The point of an apparently logically complex moral utterance, I think Blackburn is suggesting, is to express an attitude toward n-tuples of attitudes, where the last member of the n-tuple is non-cognitive, and earlier members might be cognitive or non-cognitive. In his expressive language, Blackburn adopts the B! and H! operators, as devices to indicate expressive illocutionary force as well as the particular attitude expressed; they function as moral pseudo-predicates would in English if expressivism were correct. 'B!' means 'Boo!,' and 'H!' means 'Hooray!' So for him, we can express an attitude toward lying, that goes in English like "Lying is wrong," like this:

B! (lying)

...and we can express an attitude about attitudes toward lying and getting your little brother to lie, that goes in English like "If lying is wrong, then getting your little brother to lie is wrong," like this:

H! (| B! (lying) | ; | B! (getting little brother to lie) |)

...where the vertical bars indicate indirect reference, and the semicolon indicates pairing. This utterance expresses the Hooray! attitude toward a certain ordered pair of non-cognitive attitudes. This notation is inadequate to the task, because it doesn't reflect the ordering of the members of the pair.[94] Blackburn has no way to distinguish between expressing the Hooray! attitude that you Boo! lying *and* that you Boo! getting little brother to lie, and expressing the Hooray! attitude that *if* you Boo! lying, that you *then also* Boo! getting little brother to lie. Though his notation is supposed to represent the latter, it clearly represents the former; this is because the semicolon is to represent "involve[ment] or coupl[ing]", [95] both of which notions are apparently conjunctive, not

[94] I draw this point from Hale 1986, pp. 72-75; he makes more or less the same argument at Hale 1993, p. 343. Blackburn 1993 contains a discussion of the latter paper, but Blackburn mainly skips defenses to the modal presentation that I discuss below.

[95] The examples, as well as the explanation of the bars and semicolon, appear in Blackburn 1984, p. 194.

conditional.[96] The original sentence, "If lying is wrong, then getting your little brother to lie is wrong," is supposed to express an attitude toward getting your little brother to lie that's conditioned by having a certain attitude toward lying. What Blackburn's interpretation gives us is, instead, an attitude toward having a pair of attitudes. The original does not express an attitude about what happens if you don't have a problem with lying, and also don't have a problem with getting your little brother to lie; Blackburn's interpretation does.

The trouble here is that it's easy to understand what it is to have a pair of attitudes, but it's not so easy to understand what it is to infer one attitude from another. The usual objects of inferences are propositions, but Blackburn takes for granted that we can infer one attitude from others just as though the attitudes were propositions. This is just what needs to be explained.[97]

For Frege, the referent of an otherwise assertive utterance in an indirect context is a thought or proposition in the abstract, not anyone's belief. Likewise, for Blackburn, the referents of the various Boo! and Hooray! utterances, embedded within an utterance expressing moral sensibility, are attitudes construed abstractly. We should not agree with Frege that there are abstract entities, thoughts, to serve as the contents of attitudes (though there are truth-conditions to do the job). But it's clear that attitudes themselves are not abstract. Blackburn's notation tries to have reference to the attitudes in the abstract, but attitudes are not abstract entities.

What expressivism needs is some way of displaying these utterances in expressive language but that also seems intuitively correct. Consider displaying Ex1 something like this (where regard is some suitable positive attitude):

> E regard: that if someone says "E contempt: that you did that" that they then say "E regard: that you be punished"

---

[96] It's possible that Blackburn has something else in mind, but his failure to define 'pairing' leaves me with no other plausible interpretive options.

[97] I draw this point from Zangwill 1992, esp. pp. 180-3, and Schueler 1988, esp. pp. 495-6.

Here's an alternative:

> E regard: that if someone has contempt: that you did that, that they then have regard: that you be punished

This seems to match Blackburn's intentions, though to be superior to his own way of carrying out those intentions. In this little logic, Blackburn's attitudes have been replaced with propositions that make reference to the attitudes; those propositions can include the ordered conditional, instead of just Blackburn's conjunctive semicolon. Further, the attitudes are not treated as though they were abstract entities.

Also, unlike with Blackburn's interpretation, we can see how moral reasoning might occur. Consider trying to figure out what attitude to have about getting your little brother to lie on the basis of the two attitudes expressed here:

> B! (lying)
>
> H! (| B! (lying) | ; | B! (getting little brother to lie) |)

The first line gives us nothing, because the second already expresses a positive attitude toward having the attitude expressed in the first line. Presumably, the attitude in the second line is sufficient to bring it about that the agent with the attitude expressed in the second line has both of the conjoined attitudes. But, since the second line expresses an attitude already causally sufficient to bring it about that the agent has the Boo! attitude toward getting little brother to lie, independently of having the Boo! attitude toward lying, there are no logical relations between the attitudes. In English, you cannot infer that "Getting your little brother to lie is wrong" from "If lying is wrong, then getting your little brother to lie is wrong" in the absence of an affirmation of the antecedent. But that's not true in Blackburn's notation.

Contrast that with the present notation:

> E contempt: that someone lies
>
> E regard: that, if someone has contempt: that someone lies, then, she has contempt: that someone gets his little brother to lie.

The second line expresses an attitude about what other attitudes someone should have, given that she has the attitude expressed in the first line. So someone who can sincerely perform both of these expressive acts should be caused, in a rationalizing way, to have the consequent attitude in the second line.

So the correct translation of Ex1 should be clear. But what about:

Ex2) If your action was murder, then your action was wrong.

That isn't too hard. We can notate this as:

> E regard: that if someone says " ⊢ that your action was murder" that they then say "E contempt: that you did that action"

Here's an alternative:

> E regard: that if someone has belief: that your did action was murder, that they then have contempt: that you did that action

In Ex1, the antecedent is normative, and so is the consequent. So we can interpret the utterance, expressivistically, as expressing an attitude toward having one attitude in case you have another one. In Ex2, the antecedent is non-normative, and the consequent normative. So we can interpret the utterance, expressivistically, as expressing an attitude toward having a certain attitude in case you have a certain belief. Ex1 and Ex2, then, both relate attitudes to one another; the first relates two desires, the second a belief to a desire. But the following will be untranslatable:

Ex3) If your action was wrong, then you will be punished.

Here, the antecedent is normative, but the consequent non-normative. Unless the speaker is trying to suggest that whether or not you will be punished depends on how she feels about your action, there's no way to interpret this utterance in the expressivist notation I'm developing. The problem is that this sentence can't reasonably be construed as relating two attitudes. It takes something the expressivist construes as an attitude, and relates it to an independent fact. For further examples, consider:

-No good deed goes unpunished.

-The most immoral politician is the one with the best chance of getting elected.

Surely my high regard for your deed does not trigger punishment of you, and my low regard for a politician's character does not increase his chances of election. So the (rather silly) interpretation that the expressivist could offer obviously fails.

Ex3 is a moralistic *explanation*.[98] These explanations at least appear to appeal to moral facts. There may be no such facts to which to appeal, in which case every such explanation is false. But if our moral discourse never appeals to moral facts, as it would not if expressivism were correct, there would be no such thing as moralistic explanations, whether true or false. Yet there are.

A first move that the expressivist might try is to adopt a local error theory of moral explanations: most of our moral discourse is expressive and fully successful, but all of our moralistic explanations are cognitivist and false. Such a move would be implausibly *ad hoc*: what's the expressivist explanation for our indulging in moral explanations, if nowhere else in our moral discourse do we speak as though there were moral facts?

Blackburn tries three maneuvers to deal with moralistic explanations; I'll discuss two of them in this section and the third in the next. The first is to suggest that "the explanations are elliptical. Someone's citing injustice as the cause of revolution might be adverting to the population's *perception* of injustice, or belief that they are victims of injustice." (Blackburn 1990, pp. 204-5) Since, according to expressivism, there is no injustice to perceive and no beliefs about injustice, we need to reconstrue Blackburn's interpretation. For Blackburn, injustice played no causal role in the revolution; negative attitudes toward the rulers by the revolting populace must be the explanation intended by the speaker. This maneuver can be defeated by sheer stubbornness. If a speaker insists

[98] Gilbert Harman discusses moral explanations in Harman 1977, especially pp. 3-26; this led into a debate with Nicholas Sturgeon that includes Sturgeon 1984, Harman 1986, Sturgeon 1986a, and which led into Brighouse 1990, Blackburn 1990, and Sturgeon 1991. I've especially drawn from Brighouse's paper for this discussion of moral explanations.

that he was, in fact, trying to explain the revolution with reference to the injustice of the rulers, and not with reference to anyone's attitudes about anything, then the expressivist can no longer offer her reinterpretation. In some Christian and Marxist traditions, history is regarded as teleological, with the moral goal of a certain kind of society drawing events toward it without anyone so intending. Likewise, Confucian discussions of the Mandate of Heaven might not be best understood as indirect discussions of peoples' attitudes, but as discussions of the causal power of a real moral force. The expressivist can't correctly interpret these traditions if she refuses to accept that they are making the claims they make, and those claims turn on the explanatory power of morality.

Blackburn admits that the first maneuver will often fail. For instance, the decency and humanity with which a child is raised will have an effect on that child's dispositions and well-being, independent of the child's beliefs about the decency and humanity of her upbringing. Thus Blackburn introduces:

> ...the second projectivist line, which is that the explanation points downward to the properties upon which the moral verdict depends. According to me, an upbringing is decent and humane *in virtue of* other features — meeting the child's needs, engaging with its attempts at action and communication, and so on — and I may simply point toward those other, causally powerful properties by using the moral predicate. (*ibid*)

What are 'the properties upon which the moral verdict depends?' The moral verdict is one of the moral attitudes, and those attitudes are brought about by features of the world. For instance, the attitude "Hooray! That you bring up your child like this" is caused by the facts that you meet your child's needs, engage with its attempts at action and communication, and so on.

This interpretation will be hard for the expressivist to sustain. Imagine that I become aware that you bring up your child in such a way that her needs are met. Alternately, imagine that I become aware that you bring up your child in such a way that you engage with her attempts at action and communication. My reaction in either case will be the same: I'll regard your child-rearing practices as decent and humane. But why

is that? For the expressivist, there is no such property, decency or humanity, that your different child-rearing practices instantiate. So what makes both your meeting of your child's needs, and your engagement with her attempts at action, instances of decency or humanity? Nothing holds the two physically distinct practices together morally. The expressivist must posit two psychological laws: seeing a child's needs met will generate a certain attitudinal response, and seeing a parent engage with a child will generate the same attitudinal response, even though the two stimuli are in no way similar except that they trigger the same response. The expressivist then owes us an explanation of this phenomenon: why the same response to different stimuli? The realist, on the other hand, can claim that these are different instances of the same stimuli: decency, or humanity. Each instance of meeting a child's needs is also an instance of decency or humanity. The realist then has an explanation for the similar responses.

But even if the expressivist could deal with moralistic explanations, she has not dealt with every conditional with a normative antecedent and a non-normative consequent. Consider another conditional: "If Bob is rational and carries an umbrella, then Bob believes that it's going to rain." Action explanations and content attributions are all carried out against the normative background of rationality; this is a consequence of attitudinal holism. If Bob is a lunatic and carries an umbrella, then we don't know anything at all about his beliefs; he may be carrying the umbrella to protect himself from Napoleon, the Emperor of the Hippopotami. We can make content attributions on the basis of observable behavior only in light of the relations of rationalization that lie between the attributed content and the observable behavior. We attribute to people belief in the content of their assertions, for instance, but we would stop if we found out that they were compulsive liars or otherwise irrational about their speech.

Content attributions made on the basis of observable behavior are not explained by the behavior. So an expressivist account of moral explanations wouldn't help with utterances relating content attributions to observable behavior relative to the normative

property of rationality. There are more conditionals with normative antecedents and non-normative consequents than there are conditionals that are normative explanations. Blackburn's first attempt to solve the Frege-Geach problem appears to fail when confronted by some moralistic conditionals.

Blackburn has offered an alternate solution to the problem in his paper "Attitudes and Contents." Here the solution is to articulate a logic for expressives that has modal roots.[99] This solution makes no reappearance in Blackburn's 1998 *Ruling Passions*, but the earlier solution does reappear. The relation between the two solutions is nebulous, and it's not clear whether Blackburn has given up the modal solution or whether he thinks that it's consistent with the first solution. In any event, the modal solution is not successful.

The solution begins with standard definitions of models and model sets, but moves on to allow the model sets to include special expressive sentences with H! and T! operators. The H! operator is familiar from the H! and B! operators earlier. The T! operator is to express tolerance of the propositional content of the expressive. There is no B! operator. The general idea is that the H! operator specifies a propositional content that is true in every ideal world, while the T! operator specifies a propositional content that is true in some ideal world. H! thus corresponds to expressive necessity, T! to expressive possibility. Hence:

$\sim H!(p) \leftrightarrow T!\sim(p)$

$\sim T!(p) \leftrightarrow H!\sim(p)$

$\sim H!\sim(p) \leftrightarrow T!(p)$

$\sim T!\sim(p) \leftrightarrow H!(p)$

H! and T! sentences, of course, aren't susceptible to being true or false. Thus the negation signs that appear before the H! and T! markers are not the truth-functional negation. It's unclear what they might be; an expressive act can't be truth-functionally negated, but

[99] He draws on Hintikka 1969a.

there's no other kind of negation in the offing. We negate real moral utterances all the time, but it's not possible to negate expressive speech acts. What it is to negate one of the H! or T! sentences is what needs to be explained.

This may be the most basic version of the Frege-Geach problem. Our real moral discourse employs negation over the entire contents of speech acts. But expressive discourse can't. The attitude presented by an expressive speech act can't be negated in a speech act. Consider the example of expressing a positive attitude toward the Yankees; this might take the form of "Go Yankees!" What would it be to perform an expressive speech act in which the positive attitude itself is negated? It's easy to express a different, conflicting attitude ("Boo Yankees!"). But how would one negate the positive attitude? If it appears negated, it must not exist; if it doesn't exist, then it can't be sincerely expressed. So far as I know, it's not possible to express the negation of an attitude. It's not clear what the negation of an attitude (as distinct from the absence of the attitude) would even be.

The use of negation in a discourse is the most obvious sign that that discourse is intended to be truth-apt.[100] It's through negation that falsehood, and hence truth, most obviously enters a discourse. So the presence of negation in moral discourse is the most obvious sign that moral discourse is truth-apt and that expressivism is false. Any expressivist re-interpretation of our moral discourse must either misinterpret that discourse, by leaving out negation of the entire contents of speech acts, or else introduce negation and, along with it, truth.

Though they can't be true or false, H! and T! sentences can be *satisfied*, by way of taking contents that obtain. If *p* is true in a world, then H!(*p*) is satisfied in the same world. Consistency is a relation between satisfaction conditions. H!(*p*) is inconsistent with H!~(*p*), because the satisfaction of an H! sentence in a world requires the obtaining

[100] Thanks to Josh Dever for helping me get clearer on the relation between negation and truth-aptness. I should also mention that Dever's introduction of material on minimalism and expressivism (the debate starting with Smith, Michael, 1994) was enormously helpful to me when it came to understanding Blackburn later on.

of its content in that world, but T!(p) is consistent with T!~(p), as the satisfaction of a T! sentence in a world doesn't require the obtaining of its content in that world, but only in some of the next successive realizations of the ideal relative to that world.

But now how does this help with moral conditionals, especially with moral antecedents and non-moral consequents? The normative and non-normative sentences can be connected by the conditional; for instance, $p \rightarrow$ H!(q) might represent "If you committed a murder, then you should be punished," with p representing "you committed a murder" and q representing "you are punished." The conditional, then, will say that, if we're in a world in which you committed a murder, then we're in a world in which Hooray! You being punished. But just what it is to *imply* "Hooray! something" is what needs to be explained.

Further, there's no progress on the most problematic conditionals, ones with normative antecedents and non-normative consequents. H!(p) $\rightarrow q$ makes no sense, as it appears to say that, if I express a positive attitude toward some fact, then some other fact obtains. That is, as a rule, not true, and not what the speaker wants to say. The problem, as before, is not solved by translation schemes. What it is for "Hooray! something" to imply something else (other than another expressive) is what needs to be explained.

Blackburn explains how this is supposed to help: "...endorsing the involvement [e.g., (~A v C), as a reading of ($A \rightarrow C$)] is tying oneself to the tree: in other words, tying oneself to restricting admissible alternatives to those in which ~A, and those in which C. You have one or the other." (Blackburn 1988b, p. 197) But to what, exactly, is the speaker committing herself? Assume that a speaker says that H!(p) $\rightarrow q$. She has said that q is true under a certain circumstance. But what circumstance is that? It can't be that H!(p) is satisfied; if that were the case then:

[H!(p) $\rightarrow q$] $\leftrightarrow$ [$p \rightarrow q$]

...which is, at best, obviously false; at worst, we really don't know what to do with the entire sentence as it truth-functionally relates something without a truth-value to things

with truth-values. The only thing that *q* can depend on is the attitude expressed by H!(*p*). But it's obvious that, if I say that no good deed goes unpunished, I am not saying that my regard for a deed causes it to bring punishment.

Blackburn can't account for certain of our moral utterances: conditionals with normative antecedents and non-normative consequents. The point is not that he must claim that they are all false; he can't claim that. The point is that he must systematically misinterpret them. His two solutions to the problem are notational variations on a theme, and neither of them can explain how it is that I can try to explain an event with reference to a normative fact if nothing in my language suggests the existence of normative facts. Cognitivism takes a large inductive lead over expressivism, because cognitivism can correctly interpret these utterances that, correctly or incorrectly, we make with some frequency.

4.4 Expressivism and Truth

The second embedding problem for the expressivist deals with moral utterances appearing in sentences in which they precede "...is true," or follow "I believe that..." or "It's true that...." Of course, these aren't the only such sentences, but they're the core sample. When we utter sentences like these, we seem to be asserting that various propositions are true, or that they're the contents of our beliefs. But for expressivism, our moral utterances don't present any propositional contents as true; nor do we have moral beliefs. So, for expressivism, it seems that, "It's true that murder is wrong" is false, even though the expressivist might agree that murder is wrong. Further, even an expressivist who sincerely agrees that murder is wrong would not be able to sincerely say, "I believe that murder is wrong," since, of course, if she's right, then she has no moral beliefs. (Thus the expressivist is in the odd position of having to say the following: "Murder is wrong. But what I've just said is not true and I don't believe it.") Since "It's true that..." and "I believe that..." utterances with complete moral sentences replacing the ellipses play a major role in real-life moral discourse, it will turn out, on expressivism, that quite a lot of

our real-life moral discourse will be false. But a major point of expressivism was to save our moral discourse from the threat of moral scepticism; unless these consequences can be avoided, expressivism has failed to achieve its objective. But further, the fact that these moral utterances exist in contradiction to expressivist notions about the point of our moral discourse provides strong evidence that expressivism is false.

Blackburn's strategy for rejecting the apparent consequence involves trying to get moral utterances to be truth-apt, and to present the contents of beliefs, while denying that there's anything cognitive or representational about truth-aptness or beliefs. The strategy has actually shown up twice before, in sections 4.2.2 and 4.3. In 4.2.2, I discussed the metaphysical argument that expressivism can account for the supervenience of the moral on the non-moral, while realism can't. One of the problems with the metaphysical argument is just trying to phrase it: if there are no moral properties, then it's hard to explain how they supervene on anything. In 4.2.2, I gave the expressivist some properties — of causing moralistic responses in people — that went proxy for real moral properties. But I also pointed out that the expressivist might try to introduce moral properties, while insisting that the expressivist's moral properties don't violate the ban on moral properties, though the realist's moral properties do. For the expressivist to pursue that strategy is for the expressivist to introduce something corresponding to moral predicates, and hence to do a large part of the work for the introduction of moral propositions; if the expressivist could carry off such a project in a non-realist way, then she is closer to solving the second embedding problem.

In 4.3, I noted that Blackburn presents three maneuvers designed to help deal with the phenomenon of moral explanation. While two of them were found wanting in 4.3, one of them remains to be dealt with. The remaining maneuver is the attempt to introduce moral explanations that don't violate the ban on using moral properties to explain things.

The original Frege-Geach problem would be solved as a side-effect of some ways of solving the second embedding problem. The Frege-Geach problem emerges from the

fact that moral utterances sometimes appear in contexts that require truth-apt utterances, but expressive speech acts aren't truth-apt. But if they were truth-apt after all, then the Frege-Geach problem would disappear.

The maneuver used to solve these problems (and, by extension, the Frege-Geach problem) is minimalism about truth, belief, and properties.[101] I'll begin by addressing minimalism about properties, since this allows me to tie up the loose ends I've just mentioned and leads in to the deeper issues about truth and belief. The reason to introduce minimalism about properties is to solve the problem about moral explanations I discussed in the last section, as well as to make sure that we can express the supervenience of the moral on the non-moral. On this view, *F-ness* is a property just in case a sentence of the form "*Fa*" is meaningful, for some suitable *a*.

According to Blackburn, should the first two solutions to the problem of moral explanation fail, the expressivist can always just introduce moral properties with explanatory power that are nevertheless not the properties that the moral anti-realist was concerned to deny:

> It will not be obvious that this position is available to the projectivist. But here is a sketch of the way it might be. The first part is to establish our right to talk of the moral feature or property. Now, if the projectivist adopts quasi-realism, he ends up friendly to moral predicates and moral truth. He can say with everyone else that various social arrangements are unjust, and that it is true that this is so. Once this is said, no further theoretical risks are taken by saying that injustice is a *feature* of such arrangements, or a quality that they possess and that others do not. The first step, in other words, is to allow propositional [i.e., assertive] forms of discourse, and once that is done we have the moral predicate, and features are simply abstractions from predicates. (Blackburn 1990, p. 206)

---

[101] The combination of minimalism (or related deflationary views) with expressivism, despite its contradictory nature, is actually a classic view. Perhaps the first to combine expressivism with minimalism was Ayer himself. Seminal for contemporary discussions of the problems with trying to combine expressivism with minimalism is Boghossian 1990. Several philosophers agree with Blackburn that minimalism and expressivism make a nice pair; see Stoljar 1993, Smith, Michael, 1994a, b, Jackson, Oppy, and Smith 1994. For the view that minimalism demotivates expressivism, see Divers and Miller 1994 and Horwich 1994. I draw heavily on the latter papers. I also ignore my discussion of minimalism from chapter one, since focusing on the errors of minimalism would distract from trying to understand Blackburn.

Blackburn speaks of quasi-realism, his famous project that I haven't yet mentioned. According to the expressivist, our moral utterances have as their illocutionary point to express desires. Yet their overt linguistic shape is undeniably assertive, not expressive. So the expressivist must explain why, though there's nothing assertive about our moral discourse, nevertheless, there's nothing expressive about the appearance of our moral discourse. Quasi-realism is Blackburn's attempt to do this job of explaining.

Section 4.3 was partly about one version, the translation version, of quasi-realism. On translational quasi-realism, the quasi-realist provides translations into an overtly expressive language of our apparently assertive moral utterances. Then she explains why expressive moral language is, so to speak, our first moral language, and why we always speak in a second, apparently assertive, language. The translation schemes between real moral language and expressive language are the first thing the translational quasi-realist must generate. To explain why we dress up our ejaculations in assertive language, the expressivist must first explain which ejaculations we have dressed up, and how; she must systematically relate our actual moral language to some moral language that wears its expressive nature "on its sleeve." In 4.3, I showed how an expressivist could translate some of our moral utterances into expressivistic language, but that she could not translate our moral explanations into expressivist language; I conclude that translational quasi-realism fails. Stating what non-cognitive attitude we express in giving moral explanations is part of translational quasi-realism. Unless that sub-project is complete, translational quasi-realism can't be appealed to. Blackburn's third way of dealing with moral explanations begins by appealing to his completed project (that the expressivist "adopts quasi-realism" is sufficient for her to "become friendly to moral predicates and moral truth"); that's the project that he's working on at the moment. He can't further his project by appealing to it.

I want to look at how that third way was to continue once quasi-realism allowed for the introduction of moral properties. Expressivism has more than the semantic

problem of securing reference to moral properties; it also has the metaphysical problem of accounting for the explanatory powers of those properties. If the expressivist secures reference to moral properties, but in such a way as to guarantee that any moral properties to which we can refer has no explanatory power, then it will turn out that all of our moral explanations were false; again, that we give such explanations will exceed the explanatory power of expressivism, and expressivism will have to yield before moral scepticism.

Blackburn explains that moral properties might have causal power because an account, similar to Davidson's and Lewis's mental-physical token-identity theories, or Jackson and Pettit's programme explanation theory, might hold for moral properties. Each instance of a moral property might also be an instance of some physical property with causal powers, and that would account for the causal powers of the instance of the moral property.

The trouble with this extension of the solution is that it's substantive and realist, not minimalist. Consider it in relation to the analogous Davidsonian account of mental causation. For Davidson, each mental event is also a physical event. This is obviously a realist account of the mental; for on the account, the mental is the physical, and we're realists about the physical. Likewise for Blackburn's account of moral properties. If the moral is the physical, then, unless we're anti-realists about the physical, then we're realists about the moral. Blackburn can introduce moral properties into moral explanations only at the expense of affirming realism about moral properties.

The translational method is not the only version of quasi-realism. There is also the minimalist version. On minimalism about truth and belief, "*p*" is truth-apt just in case sentences of the following forms are meaningful: "'*p*' is true," or "It's true that *p*"; "*p*" presents a belief just in case a sentence of the following form is meaningful: "I believe that *p*."

On this version, there's no contradiction between a language "wearing its expressive force on its sleeve" and being truth-apt and presenting the contents of belief. Every utterance that appears to assert the obtaining of content will do so; every utterance that appears to present the content of a belief will do so. Minimalism removes any appearance-reality gap when it comes to assertiveness. Looking like an assertion, for minimalism, is sufficient for being an assertion.

However, minimalism implies the first claim of realism, cognitivism: some moral claims are truth-apt. That's why Blackburn adopts minimalism to save expressivism from the second embedding problem. If we accept minimalism, then we have a neat explanation for the truth-aptness of moral utterances, and that truth-aptness would solve the second embedding problem. However, the truth-aptness of moral utterances is precisely what cognitivism claims exists, and what any non-cognitivism must deny. Blackburn seems to be explaining phenomena that he should be denying. Blackburn tries to avoid this implication in two ways. Here's the first:

> Subtlety with the concept of belief, or with the concept of truth or of fact, may enable the expressivist to soften this opposition [between realism and expressivism]. Theory may enable us to understand how a commitment with its center in the expression of subjective determinations of the mind can also function as expressing belief, or be capable of sustaining the truth predicate — properly called 'true' or 'false'.... It means separating truth (in this application at least) from 'represents' and its allies, but nobody has ever pointed out the harm in that. (Blackburn 1993, p. 185)

The parenthetical phrase makes the passage a little hard. There are two possibilities. On one possibility, Blackburn adopts general minimalism, minimalism about all regions of discourse, and holds that the minimalist concept of truth is the one and only concept of truth. In that case, his view entails cognitivism. The cognitivism of the minimalist is thin, in that truth has been separated from notions it had been thought connected with, such as representation or correspondence. But nevertheless, if minimalism is the correct theory of truth, then all that the realist needed to show in order to defend her claim that moral utterances are truth-apt is that they are minimalistically truth-apt; since it's obvious that

our moral utterances are minimalistically truth-apt, cognitivism will be a truism. Minimalism, so far from giving expressivism greater explanatory power, makes cognitivism trivially true (and hence any theory, such as expressivism, that denies it, trivially false).

However, it may be that truth is to be separated from representation or correspondence only in some regions of discourse. All truth concepts would share the commitments of minimalism about truth. But some truth concepts might have substance, and different truth concepts might have different substance. It might be that the truth concept that applies to our non-normative discourse is the correspondence notion of truth, while the truth concept that applies to our normative discourse has more to do with acceptability at the end of normative reflection, or in reflective equilibrium. In that case, since Blackburn isn't applying a cognitive notion of truth to normative discourse, and is not applying to normative discourse the same notion of truth he applies to non-normative discourse, his minimalism will have no implications about the cognitive content of normative utterances. In that case, Blackburn's minimalism would allow him to have his truth-aptness, without accepting anything disreputably cognitive.

Blackburn gives no reason to believe that we do in fact mean to say something different about a moral utterance when we embed it in "It's true that..." than when we embed a non-moral proposition in the same context.[102] But more important, the solution is problematic in other ways. Consider the Frege-Geach problem. Translational quasi-realism failed to solve the Frege-Geach problem, but minimalist quasi-realism held out hope. Even if minimalism plus a diversity of truth concepts could solve the second embedding problem, it could not solve the first. A valid argument is truth-preserving. But if the premises of an argument had different sorts of truth, what sort of truth would be

[102] The fact that he offers no such reason encourages us in the belief that he intends to adopt minimalism across the board, but his parenthetical remark would then be misleading. There's really no telling exactly where Blackburn intends to apply his minimalism.

preserved in the conclusion? Would it be substantive, cognitive, representative truth? Or minimal, expressive, non-representative truth? Consider this example:

> You are a captain.
> A captain should go down with his ship.
> Thus, you should go down with your ship.

The argument appears to be valid. But the truth concepts to be applied to the two premises are different, since one of the premises is non-normative and the other is normative. So which kind of truth is preserved through to the conclusion? Since the conclusion is normative, presumably we're to apply to it the notion of truth that applied to the second premise. But this notion of truth is non-cognitivist and non-correspondentist; it has to do, perhaps, with what we would accept at the end of moral reflection. What is the reason for believing that the first premise, which was true in a cognitive way, would have implications about what we would accept at the end of moral reflection? Also, consider this example:

> You should go down with your ship.
> Only captains should go down with their ships.
> Thus, you are a captain.

Again, the argument is valid. The truth of the premises should be preserved through to the conclusion. But while both premises are normative, the conclusion is not. Somehow, the normative notion of truth has dropped out of this inference, and been replaced with the cognitive notion of truth, even though truth was preserved.

I've looked at two versions of quasi-realism, translational and minimalist. Translational quasi-realism failed to even get started, since it couldn't offer translations for some of our moral utterances. Minimalist quasi-realism fails for one of two reasons. First, if minimalism is adopted across the board, then cognitivism is a direct implication, and expressivism is false. Second, if minimalism is adopted only for moral utterances or as a basis for assigning different notions of truth to different regions of discourse, then the original Frege-Geach problem, that various apparently valid inferences would go

invalid, would remain unsolved; also, the second reading of minimalism has no buttressing argument.

I want to continue to consider the prospects of general minimalism as a defense of expressivism, since we can gain some insight into Blackburn's proposal by looking into this line of argument; this is also where I'll look at Blackburn's second attempt to avoid the cognitivist implication of his minimalist maneuver. On general minimalism, expressivism appears to have been a waste of time. Realism consists partly in cognitivism, the claim that our moral utterances are truth-apt. Expressivism was supposed to be an alternative to realism because it's an alternative to cognitivism. But by adopting general minimalism, the expressivist accepts cognitivism. Why the detour? If we were going to end up as cognitivists anyway, and we were going to be forced to by evidence that was available at the outset, why waste our time with expressivism?[103] Blackburn replies,

> But in fact this is no objection, and there is no tail-biting. We must remember Wittgenstein's dismissive attitude to invocations of truth and representation when he is dealing with the kinds of commitment that interested him. Just *because* of minimalism about truth and representation, there is no objection to tossing them in for free, at the end. But the commitments must *first* be understood in other terms. (Blackburn 1998, p. 80; second emphasis mine)

This contrast between the beginning and end of inquiry is a frequent theme in Blackburn; here it shows up again:

> ...by now many readers will be wondering what remains that is distinctive of projectivism. In the old days (they will complain) it was easy: we knew what emotivists and prescriptivists stood for. But this new, conciliatory position is harder to pin down, harder to recognize as a position. My view has always been that it is a question of explanation, of 'placing' our propensity for ethics within a satisfactory naturalistic view of ourselves. I distinguish between the ingredients with which you start, and what you can legitimately end up saying as you finish. To place ethics, I deny that we can help ourselves to moral features and explanations from the beginning. We have to see them as constructions, or, as I call it, projections, regarding ourselves as in the first instance as devices sensitive

[103] Sturgeon addresses the issue of whether the expressivist and the realist should accept this position and stop arguing in Sturgeon 1986b.

> only to natural facts and producing only explicable reactions to them. (Blackburn 1990, pp. 207-8)

What I think this remark might suggest is that Blackburn is in fact *pursuing a project*, not *stating a theory*. When it comes to the truth of a theory, it doesn't matter in what order you say things. But when it comes to pursuing a project, you must go in the right order.

I'll call this project explanatory quasi-realism, since I think it's a third attempt to escape the embedding problems. While we are to "construct" moral properties and explanations, and explain why we speak as though there were such things, we have to construct them out of something and explain them on the basis of something. We are to construct them out of "natural" facts, and "explicable" responses to those facts, because these are available to us "in the first instance." But first instance of what sequence? By defining alternative sequences in which we progress, constructively, from expressivist reality to cognitivist appearance, we define alternative versions of explanatory quasi-realism.

Explanatory quasi-realism may be the larger project within which translational quasi-realism would find life. Explanatory quasi-realism will give some constructive sequence. At the beginning of the sequence, there will be nothing moralistic; at the end will be our seemingly cognitivist moral discourse. To give intuitive sense to the beginning of the sequence, we can use a moral language that displays its expressive nature on its sleeve. The construction will be the explanation of how we got to our real moral discourse from the original, expressive discourse. Translation, then, will play at least a pedagogical or presentational role in explanatory quasi-realism, because the translations give us access to the explanatory base.

It may be that the construction happened across historical time. Perhaps Blackburn thinks that there was once a primal moral language that was overtly expressive, and that that language shifted into our apparently assertive moral language. He certainly hasn't said anything that would explain that shift, and so hasn't said anything

relevant to reconstructing any such shift. Maybe the sequence is psychological or biographical. Perhaps Blackburn thinks that we construct our moral language on the basis of desires in much the way the positivists thought that we construct our scientific language on the basis of scraps of sensation. If that's the goal, then he's started the reconstruction. The first step in such a reduction of scientific language to sensation language would be to generate a sensation language into which we could translate scientific language; the second step would be to prove that the scientific language was actually constructed out of the sensation language to which we're trying to reduce it. Blackburn has tried to generate some parts of the corresponding expressive language, though he's never said anything that would encourage us to think that the expressive language he created provides the basis for our moral language. So again, we have little guidance in guessing what the project might be.

A third option would have it that the sequence is from reality to appearance. Our real moral discourse is expressive; it only appears to be assertive. Maybe what's to be explained is why there is something in the world that takes this particular appearance. In reality, there are only desires, and expressions of those desires, and actions to satisfy those desires. The desires appear in the guise of beliefs, the expressions appear in the guise of assertions, and the actions to satisfy the desires appear in the guise of behavior subject to moral considerations (of which there are none in reality). The difference between expressivism and scepticism is that the expressivist will try to "preserve the appearances" in the sense that she won't try to argue us out of accepting those appearances (in some sense of "accept"). But if that's the sequence, then the reality seems inexplicable to expressivism. We treat moral utterances as truth-apt, but they aren't really. Why would we treat moral utterances as truth-apt, if they weren't? And why wouldn't we be in error to do so? Blackburn has created expressivist languages for moral discourse in his effort to explain what we really mean by our apparently assertive moral discourse. Why don't we speak one of those? Expressivism is a non-starter until it can explain why

our cognitive-seeming moral discourse looks so very little like the non-cognitive reality allegedly underlying it.

5 Moral Realism and Moral Scepticism

5.0 INTRODUCTION

This chapter is the culmination of my defense of moral realism. In it, I offer a new version of moral realism called Hermeneutical Moral Realism, and show that this theory can explain what expressivists cannot, offer an appealing approach to moral principles, and withstand very sophisticated sceptical attacks.

In the first section, I begin to place my theory within an historical context by discussing what's known as New Wave Moral Realism. New Wave Moral Realism is the nearest logical relative to Hermeneutical Moral Realism; it's my Kantian approach's Aristotelian cousin. The New Wave theory is designed to respond to Moore's Open Question argument, and I discuss the dialectic between Moore's concern and the New Wave.

In the second section, I offer my own version of moral realism. I stress the application of externalism and holism to the hermeneutics, and hence the nature, of moral beliefs. The fact that moral beliefs' contents are their truth-conditions undermines moral scepticism. The holistic nature of moral beliefs underwrites an approach to the nature and importance of moral principles that is modestly particularistic, and yet recognizes that the employment of moral principles is essential to most moral reasoning. Further, inter-attitudinal holism accounts for the relationship between moral beliefs and moral action more satisfyingly than alternative accounts.

In the final section, I reply to the moral sceptic. First, I deal with J. L. Mackie's standard arguments. Then, I turn to a later development offered in reply to the New Wave, the Moral Twin Earth argument. I contend that the Moral Twin Earth argument succeeds against the New Wave, but, because the New Wave was based on an erroneous

version of externalism, Moral Twin Earth is not a counterexample to Hermeneutical Moral Realism.

5.1 NEW WAVE MORAL REALISM

The Moral Twin Earth Argument is directed at what its inventors call New Wave Moral Realism. New Wave Moral Realism is a version of moral realism developed in response to Moore's Open Question Argument, itself largely responsible for the demise of moral realism earlier in the twentieth century. The response was made possible by developments in philosophy of mind, language, and science.

I begin with the Open Question Argument. For Moore, "...propositions about the good are all of them synthetic and never analytic." (Moore 1903/1993, p. 58) Why would that matter? That there can be an analytic proposition about x is the mark of x's complexity; that which is analyzed is a complex of parts into which it can be analyzed. So goodness, as shown by its unanalyzability, is simple. And that matters because "Definitions... which describe the real nature of the object or notion denoted by a word, and which do not merely tell us what the word is used to mean, are only possible when the object or notion in question is something complex." (*ibid.* p. 59) If goodness were simple, then it would be impossible to describe its real nature. Why should we believe that goodness is simple? "...whatever definition [of goodness] may be offered, it may be always asked, with significance, of the complex so defined, whether it is itself good." (*ibid.* p. 67) If we were to assert that, analytically or definitionally, goodness is that which is conducive to human well-being, and we were to agree that something in particular is conducive to human well-being, it would still make sense to ask whether it is good. However small the step from agreeing that something is conducive to human well-being to agreeing that it is good, there is a step, and that disproves the analyticity of the claim that goodness is that which is conducive to human well-being. We could agree with the claim, of course, we just couldn't hold it to be true analytically.

Since goodness is indefinable, it isn't possible to state its nature. Were its nature something natural, then it seems that it would be possible to state it; hence, it isn't something natural, it's something non-natural. With the demise of such Platonic pleasantries as non-natural properties, moral realism itself collapsed and anti-realist ethical systems like emotivism came to the fore.

The revival of Moral Realism that I want to address, New Wave Moral Realism, takes its lead from alternatives to Moore's semantic presuppositions. For Moore, the fact that we can't *a priori* close the open question shows that the *definiens* doesn't define goodness: meanings are mental or Platonic entities, and their sameness is apparent. Brink notes that the Open Question Argument

> ...appears to be a consequence of the traditional theory of meaning according to which the meaning of a term is the set of properties that any speaker competent with the term associates with it. It seems hard to believe, on this theory of meaning, that one could doubt that two terms were synonymous [as one can with 'good' and any proposed *definiens*] if the properties one associated with them were the same. (Brink 1989, p. 153)

Content internalism, one of my main targets in the first two chapters of this dissertation, is the semantic context in which the Open Question Argument finds life. Of course, it would be denying the antecedent to conclude that moral realism must be true because one argument against it is rooted in bad semantics. But better semantics opens up a space in which new versions of moral realism can be tried.

Horgan and Timmons, in setting up the target of their Moral Twin Earth Argument, gave several[104] developments in mid-century philosophy that undergirded the New Wave. First was "rejection of a synonymy criterion of property identity." (Horgan and Timmons 1990, p. 451) That criterion was active in the Open Question Argument: failure of synonymy between *definiens* and 'good' was what was to show failure of identity. Thanks to Putnam[105] and others, properties may now be identified *synthetically*.

---

[104] Six, to be precise; I'll ignore the fifth, since it seems to be a rephrasing of the first and second.
[105] See, obviously, Putnam 1975a.

Second was the notion of rigid designation, a causal, externalistic conception of reference.[106] With the older approach to reference, reference was a matter of a match of some kind between a meaning — a cluster of properties subjectively associated with a word — and referents. After the Kripke-Putnam semantic revolution, many referring terms were held to be rigid designators, terms that referred to some object or property regardless of subjective associations. If 'good' were a rigid designator, then it would refer to whatever property it referred to quite regardless of our inability to identify that property. Of special relevance is the notion of rigid designation of a natural kind. Not only objects can be rigidly designated: properties can, too, if they have appropriate causal powers. Such properties are the natural kinds, and they are the properties that science will accept as it nears "the end of inquiry." If goodness were such a property, then it could be a naturalistic property despite having no known definition, just as many natural kind terms have no known definitions prior to the discovery of their essences by some appropriate science; e.g., chemistry's discovery of the essence of water, physics' discovery of the essence of light, and so forth. Moreover, if goodness were a natural kind, then there could be a pleasant and perhaps instructive connection between ethics and biology, since biology is the science that identifies the essences of living phenomena, such as human beings, and ethics seems to have to do with our essence. On such a basis in quasi-Aristotelian metaphysics, there could be a revival of Aristotelian ethics.

Third, and connected with the peripatetic aspects of the last, is Putnam's functionalism about the mind.[107] For functionalism, what it is to be a mental state of a certain kind is to be a token of some (paradigmatically biological) kind, tokens of which typically have certain causal relationships. These typical causal relationships determine what sort of mental state is in question, because a typical causal relationship, if selected for, is a functional state, and mental states are understood functionally. What has this to

[106] See, obviously, Kripke 1972.
[107] On which see Putnam, 1960, and Putnam, 1973, among a great many works by Putnam and others.

do with the New Wave?[108] Functionalism provided a paradigm for the naturalistic treatment of concepts not susceptible to reduction to the physical. If mental concepts could be treated as concepts of functional states in a naturalistically reputable way, then perhaps moral concepts could, too. Goodness, for instance, might be connected with the flourishing — the biologically proper functioning — of a human being. A morally good person is one that functions well in certain ways. But these notions of well-being or functioning well could be given functionalistic accounts like mental concepts are in functionalism.

Fourth, it's no longer considered necessary to treat a property as type-identical to some physical property for it to be naturalistically acceptable. For functionalism, functional properties might be realized in any of many physical substrata. Token identity between mental and physical tokens is enough to preserve the naturalness of the mental. Likewise, the New Wavers will note, for the moral. As long as every instance of goodness is an instance of some physical property, goodness is natural, even if goodness is not reducible to, or type-identical with, any physical property (or narrow set of physical properties).

The last development is the development of naturalistic and coherentist approaches in epistemology. How could these be of use to the moral realist? Moral realism is plagued with the problem of how we could be in cognitive contact with the moral, if the moral is independent of the mind. But living organisms form moral beliefs as a consequence of the functioning of adapted cognitive structures; moreover, moral beliefs can cohere with one another and with relevant non-moral beliefs. If such reliable sources and unimpeachable methods can apply to morality, then there is no special problem for moral epistemology and there is new space for moral realism.

---

[108] For that matter, what has it to do with Aristotle? Aristotle has been interpreted, to my mind very plausibly, as forwarding a version of functionalism. See Nussbaum 1978 and Putnam 1973. For debate, see Burnyeat, 1992, and Nussbaum and Putnam, 1992.

Boyd and Brink are the paradigm New Wavers. I want to give a brief review of the essential elements of Boyd's views (which are more efficiently, for my purposes, presented, and massively overlap with Brink's), with special attention to the semantic background within which those views find life. Boyd gives an account of moral properties and how we may refer to them. The account of reference is solidly externalistic:

> *Roughly*, and for nondegenerate cases, a term *t* refers to a kind (property, relation, etc.) *k* just in case there exist causal mechanisms whose tendency is to bring it about, over time, that what is predicated of the term *t* will be approximately true of *k*... we may think of the properties of *k* as regulating the use of *t* (via such causal relations), and we may think of what is said using *t* as providing us with socially coordinated *epistemic access* to *k*; *t* refers to *k* (in nondegenerate cases) just in case the socially coordinated use of *t* provides significant epistemic access to *k*, and not to other kinds (properties, etc.). (Boyd 1988, p. 195)

Allow *t* to be some moral term. Then *t* is *causally* coordinated to instances of some matching moral property *p*. Notably, Boyd appeals to the anti-sceptical elements inherent in externalism when he speaks of our epistemic access to the referents of natural kind (property, and so forth) terms. Since the term has its meaning only in virtue of its causal connection to its referent, the term can't be systematically misused.

The notion of moral properties having causal powers — say, the causal power to regulate our speech — is, on face, hard to accept. Moral properties are (obviously?) irreducible to, distinct from, the physical properties that we expect to show up in serious causal laws. Boyd, however, has it that moral properties do have natural definitions; that is, they are conceptually connected with more scientifically respectable properties. He appeals to the idea of "property-cluster definitions."[109] The idea is that a term might be associated, not with some one naturalistic property, but with a set of them that, in a defeasibly nomological way, cluster together. The clustering is a consequence of some of the properties attracting others, or that which leads to the instantiation of some tending also to lead to the instantiation of others. He calls this clustering, 'homeostatic' clustering.

[109] Boyd 1988, p. 196.

In what way is this approach different from the internalist line that a term's referent is determined by the properties associated with it by the speaker? The homeostatic cluster is a natural phenomenon. The speaker need not be familiar with the distinct clustering properties, nor why (or that) they cluster. As long as the speaker's use is guided by the instantiation of the cluster of properties, the cluster is the referent of the term, quite regardless of the speaker's subjective states.

For the account to be plausible, Boyd must identify naturalistically respectable properties and explain their moral significance. The basis of his approach here is thoroughly Aristotelian. He begins with a homeostatic cluster of "important human goods, things which satisfy important human needs." (*ibid*, p. 203) The naturalism of the approach, presumably, lies in the concept of needs. Need's naturalistic credentials are certified by biology and other natural sciences (including, apparently, psychology). While Boyd doesn't say it, one can't help but think that the term regulated by this homeostatic cluster is intended to be 'happiness' — in the sense of eudaimonia. What has this naturalistically acceptable notion got to do with morality?

> Moral goodness is defined by this cluster of goods and the homeostatic mechanisms which unify them. Actions, policies, character traits, etc. are morally good to the extent to which they tend to foster the realization of these goods or to develop and sustain the homeostatic mechanisms upon which their unity depends. (*ibid*)

Happiness (in my Aristotelian usage) is a naturalistically acceptable concept. Moral goodness is defined as that which causes happiness. That which is nomologically and causally connected to the naturalistically acceptable is naturalistically acceptable. Moral goodness, then, is naturalistically acceptable.

This approach wouldn't have worked in the absence of the Kripke-Putnam semantic revolution. It's an open question whether that which is conducive to happiness is morally good. That matters a good deal on an internalistic approach to content, but much less on an externalistic approach. Internalism, plus naturalism, is lethal to moral realism.

To provide a naturalistically acceptable conception of moral goodness is automatically to provide a conception that leaves an open question.

There is another, epistemic, issue that Boyd discusses. As things stand (in my presentation), Boyd's Aristotelian ethics are just a mere assertion. How could we confirm his identification of the morally good as the happiness-conducive? Boyd's answer is that goodness's nature is discovered by just the same procedure used in the sciences for the discovery of natures. However, his view about what that procedure is, is slightly startling: "...the dialectical interplay of observations, theory, and methodology, which, according to the [scientific] realist, constitutes the *discovery* procedure for scientific inquiry *just is* the method of reflective equilibrium..." (*ibid*, p. 200) It's not obvious that reflective equilibrium has any application in the natural sciences.

Reflective equilibrium emerges in Rawls's efforts to define a method for deciding on principles of justice. Rawls has a contractarian approach to justice. But not just any contractual situation can decide on principles of justice that deserve adherence. A contract can go wrong, fail to be binding, for at least two kinds of reasons. First, the contract might be agreed to in the wrong situation. If I sign a contract under duress, it isn't binding. Second, the contract might try to oblige someone to do something impermissible. If you and I both willingly enter into a contract directing me to murder someone, then the contract isn't binding. We may assimilate the second sort of problem to the first by treating the relevant attitudes of the signatories as part of the situation. We may think that my willingness to murder someone for you is the root of the problem with the second contract. So we must design the situation in which a valid contract could be generated, making sure that the situation, both with regard to freedom to join or refrain from joining the contract, and the motivations of the signatories, can legitimate the contract. How shall we design this situation?

> By going back and forth, sometimes altering the conditions of the contractual circumstances, at others withdrawing our judgments and conforming them to principle, I assume that eventually we shall find a description of the initial

> situation that both expresses reasonable conditions and yields principles which match our considered judgments duly pruned and adjusted. This state of affairs I refer to as reflective equilibrium. It is an equilibrium because at last our principles and judgments coincide; and it is reflective since we know to what principles our judgments conform and the premises of their derivation. (Rawls 1971/1999, p. 18)

So stated, reflective equilibrium looks like the point at which our prejudices are consistent. But this passage describes only a first pass at reflective equilibrium, "narrow" reflective equilibrium. Later, Rawls proposed a "wide" conception:

> ...adopting the role of observing moral theorists, we investigate what principles people would acknowledge and accept the consequences of when they have had an opportunity to consider other plausible conceptions and to assess their supporting grounds. Taking this process to the limit, one seeks the conception, or plurality of conceptions, that would survive the rational consideration of all feasible conceptions and all reasonable arguments for them. (Rawls 1975, p. 289)

In narrow reflective equilibrium, I find an equilibrium between my own considered judgments and the general moral principles I antecedently found most plausible. In wide reflective equilibrium, I take into account general moral principles belonging to moral theories that I didn't, antecedently, find most plausible. The consideration of alternatives is to help overcome the inherent conservatism of coherentist approaches.[110]

For realism, a coherentist methodology requires an antecedent reason to believe that many of the beliefs to be rendered coherent are, in fact, true.[111] Certainly, this approach doesn't seem much like science. But that appearance is misleading. Recall the last element of Horgan and Timmons's list of recent philosophical developments undergirding New Wave moral realism. As Horgan and Timmons point out, naturalized and coherentist epistemologies are in vogue in philosophy of science. Such approaches would have it that the method of science is coherentist: the scientist seeks coherence between abstract statements of theory and particular observations. It's not at all unreasonable to believe that 'reflective equilibrium' is a moral theorist's term of art for a general coherentist approach to knowledge that applies in most areas of inquiry.

---

[110] See Daniels 1979, esp. p. 266.

[111] Rawls probably doesn't face this problem. Rawls appears to accept a coherentist theory of the truth of moral claims. Only realists with coherentist sympathies face the problem: anti-realists don't.

On the other hand, it's easy to see why science, in pursuing coherence, pursues truth. Its observations are mainly true, and its methods for relating judgments to one another, so as to capture coherence or winnow out incoherence, are unimpeachable. It's not so easy to see why coherence gives confirmation to moral views. What is the reason for suspecting that many of the moral beliefs to be rendered coherent are true? Recall that the moral properties themselves regulate our use of moral language. The New Wave must appeal to the anti-scepticism inherent in its externalistic account of moral language to show that our moral beliefs are, by and large, already in order before we begin to pursue reflective equilibrium.

Note that Rawls speaks of "derivation" of judgments from principles. Derivation is a logical relation between beliefs, not a real relation between facts. Reflective equilibrium understood as the pursuit of derivations, then, has no distinctly realist tinge to it. Brink adds that tinge when he makes this suggestion:

> ...explanatory coherence demands that we introduce more general, theoretical moral claims into our moral views in order... to unify and *explain* the more particular moral views we already hold. Coherence asks us to try to identify theoretical moral claims that will *explain* and support a number of our more firmly held moral views... (Brink 1989, p. 130; emphasis added)

For Brink, the coherentist approach in morals[112] will offer, not only derivations, but also explanations. General moral principles will explain particular moral judgments. Explanation has a worldly application that derivation lacks. Explanation turns on the capturing of real nomological connections that exist independently of theory in a way that derivation doesn't. For the New Wave moral realist, reflective equilibrium isn't a mere heuristic device. Rather, in pursuing reflective equilibrium, we pursue explanations for particular moral facts, in much the same way that the empirical scientist pursues explanations for particular natural facts in general laws.

---

[112] ...which he identifies with reflective equilibrium; Brink 1989, p. 131.

Prior to claiming that moral properties' essences are discoverable in reflective equilibrium, the New Wave had already appealed to the existence of moral properties to do more fundamental meta-ethical work. Moral properties causally govern the use of moral terms; that's why the terms have the properties as their meanings, and that's why our moral knowledge is assuredly correct in outline. The epistemic and semantic points combined show that what we will discover in reflective equilibrium governs our use, even now, of moral language. On the analogy between reflective equilibrium and scientific theory-building, that seems reasonable. It's H_2O that always governed our use of 'water;' science identifies, as it asymptotically reaches closure, just that which had already governed our use of the terms science begins with and continues to employ. Likewise, it's always (say) being conducive to human happiness that governed our use of 'good,' but only in reflective equilibrium will we be in a position to offer, synthetically, the identity of goodness.

Finally, I turn to a feature of the New Wave that Horgan and Timmons don't discuss. Brink has a distinctive view about moral motivation, one that, he thinks, is required by moral realism. In theory of moral motivation, there are alternative views about the relationship between moral facts or moral beliefs on the one hand, and moral desires or inclinations on the other. The relationships might be held to be internal, or external: facts or beliefs might be internally related to desires, or externally related. I discuss these issues more thoroughly in 5.2.4, but here I'll talk about Brink's moral externalism about moral motivation.

From my point of view, these remarks capture what's important about New Wave motivational externalism: "The externalist can claim, first, that moral considerations only *contingently* motivate or justify; second, that the motivational power or rationality of morality, whether necessary or contingent, can be known only a posteriori..." (Brink 1989, p. 42) Externalism is intended to be the negation of internalism, which has it that

moral judgment necessarily motivates, and that this motivational power is a fact in philosophical psychology knowable *a priori*.

Brink's argument for externalism consists in the claim that a certain agent, the amoralist, is conceivable. Taking Thrasymachus as a paradigm, Brink contends that we can imagine someone who makes moral judgments, but is systematically indifferent to them. Thus moral judgment does not necessarily motivate; when moral judgment motivates, the motivation is contingent on factors beyond the judgment itself, such as the specific moral psychology of the agent.

But the depth of Brink's externalism is not plain. Here is his alternative to internalist moral psychology: "We all begin with certain ends and desires; we are attached to, and have desires for, particular people, activities, and states of affairs more than others. Now these ends and desires reflect *evaluative judgments*. In the case of most things we desire, we desire them *because we think these things valuable*." (*ibid*, p. 64; emphases original) Whether this is seriously externalistic or not depends on whether Brink would admit that an amoralist agent can both think an action valuable (most valuable all things considered) and also remain indifferent to whether she performs it. It seems that he does think that evaluations typically "force" desires: "...most of our attitudes and desires *require* this sort of value-laden *explanation*." (*ibid*, p. 78, emphases mine) If evaluative judgments can explain desires, then they must have some sort of causal power over them, and there must be a nomological connection between having evaluative judgments and having "matching" desires. This seems to me to be a variety of internalism. I conclude that the New Wave has externalistic aspirations in philosophical psychology, but does not have a clear view about moral motivation.

5.2 HERMENEUTICAL MORAL REALISM

5.2.1 Outline of Hermeneutical Moral Realism

Hermeneutical Moral Realism is a version of moral realism with its argumentative basis in Davidson's approach to truth and his theories of meaning, belief, and interpretation. Here, I offer a very brief outline of the view as I will present it; in the next three sections I discuss three aspects of the view.

Moral beliefs are identified in radical interpretation. There are three aspects to the identification and interpretation of moral beliefs that are of special interest. First, the contents of moral beliefs are their truth-conditions, and interpretation takes the form, or can be understood to take the form, of Tarskian T-sentences. Davidson's anti-relativist and anti-sceptical arguments flow from these claims. It's evidence for a T-sentence of the form "'ϕ' means that ψ" that "ϕ" and "ψ" are true under the same circumstances, and too much deviation between their truth-values on various occasions demonstrates difference of meaning. Thus there is no space for moral relativism and no basis for moral scepticism.

So far, the view is standard, though more deeply defended than many other versions of moral realism. What makes it interesting and plausible is a pair of distinctive claims, both flowing from the holistic aspects of moral interpretation. The first has to do with the role of principles in moral reasoning. Theories of the place of principles can be placed on a spectrum from extreme generalism to extreme particularism. The extreme generalist wishes to deduce all moral claims from some single moral principle, plus normatively relevant non-moral facts about the situations to which the principle is to be applied. The extreme particularist holds that there is no place at all for moral principles. The view to be proposed here is a form of modest particularism; it is also a form of modest generalism. No statable principle could ever capture all moral truth. Nevertheless,

as a consequence of the interpretability of moral reasoning, there must be some fairly simple moral principles. Further, these principles must apply in most cases.

The second holistic connection is not between moral beliefs at different levels of generality, but between moral beliefs and moral motivation. The second distinctive claim has to do with moral psychology and the debate between internalists and externalists about motivation. While the labels in this area get confused, we may say that externalists about moral motivation hold that moral motivation is externally related to moral belief: it is possible to believe, say, that something would be the right thing to do, but entirely lack the desire to do it. Internalists would have it that the relation is internal: moral belief implies, or even is, moral motivation. On the basis of the holistic nature of interpretation, I argue that it counts against an interpretation if it assigns someone a moral belief without the accompanying moral motivation. However, as with all such holistic constraints, the connection between belief and motivation is defeasible: it must be present in most cases, but need not be present in all.

What's distinctive about the hermeneutical view is that it offers, in the holistic nature of interpretation, an *account* of facts generally known. It's broadly agreed that a rational agent with a certain moral belief will have the appropriate moral motivation. Rationality is typically what does the trick of accounting for the connection. But it's not clear how rationality makes beliefs cause desires. By claiming that there are inter-attitudinal connections between moral belief and moral motivation in agents that can be interpreted as possessing moral beliefs, the hermeneutical approach will explain *why* rational agents have appropriate motivation.

The core fact about moral interpretation is that it is interpretation. What's been said about interpretation, then, applies to it. However, as moral utterances are not the "plainest and methodologically most basic" utterances to interpret, we can't just apply Grace and have done with it. It's not remotely obvious that moral beliefs, even those nearest to observation, are among those beliefs whose causes are their contents; it's not

remotely obvious that moral truth-conditions have causal powers or are events. Let me offer three orthogonal distinctions among moral beliefs.

The first distinction is that between judgments and principles. As distinguished by Rawls[113], moral *judgments* are relatively low-level moral claims with narrow (perhaps unique) application, whereas more abstract *principles* apply to various contexts. Rawlsian judgments are intended to be considered convictions, but I want to let the notion of judgment range more broadly downhill to include snap judgments and moral observations. A principle is a universally quantified moral claim, one that ranges over many situations. A judgment is a moral claim that applies to only one situation. Judgments apply moral predicates to particular actions, people, institutions, and anything else that's morally evaluable. Principles apply moral predicates to all of the members of classes of morally evaluable objects. Among the moral judgments will be a special class of moral *observations*, which apply moral predicates to immediately present acts or people, on the basis of immediate experience.

The second distinction is between directives and evaluations. This is Wiggins's distinction[114] between, on the one hand, utterances about my own possible future courses of action, and utterances about all past and everyone else's actions at any time. Directives concern my beliefs about what I ought to do; evaluations concern my beliefs about what I ought to have done, what others ought to have done, and what others ought to do.

The last distinction concerns the kind of moral predicate appearing in the moral utterance: thick or thin. This is Williams's famous distinction[115] between more and less general and descriptive moral predicates. Among the thin moral predicates are terms such as 'right' and 'good,' terms of ultimate assessment that have little apparent descriptive force. If I accept someone's claim that an act would be good or right, I know very little

[113] Rawls 1971/1999, p. 18.
[114] ...drawn at Wiggins 1976, p. 95.
[115] ...drawn at Williams 1985, pp. 129-30.

about the act. Among the thick moral predicates are terms such as 'honest' and 'courageous,' terms of assessment with obvious descriptive force.

As I say, the distinctions are orthogonal. However, principles range over past and future and more broadly than the speaker, so principles are by nature neither evaluative nor directive. We then have several sorts of moral claims: evaluative judgments employing thick, and those employing thin, moral predicates; thick and thin descriptive judgments, and principles connecting non-moral with thick, non-moral with thin, and thick with thin, moral concepts.

The externalistic aspect of moral interpretation applies most directly to thick and thin evaluative judgments. The first, inter-belief, holistic aspect of moral interpretation applies most directly to principles of all kinds, and their relationships with evaluative judgments. The second, inter-attitudinal, holistic aspect of moral interpretation applies most directly to thick and thin directive judgments. Thus those sorts of judgments will be the foci of the next three sections, respectively.

5.2.2 Externalistic Aspects of Moral Interpretation

The most straightforward moral beliefs, the "most basic" in terms of interpretation, are evaluative judgments. These are the most obviously factual and mundane moral beliefs. Because they are evaluative, they don't have any obvious connection with the speaker's future actions. Because they are judgments, there is no controversy (among realists, at any rate) about whether we need them for moral reasoning.

Let me consider first the case of moral observations with thick concepts, for instance, the observation that a certain act is generous. How is such a judgment made? I suggest that we recognize generosity in much the same way that that we interpret speech (which is recognizing meanings) or identify money. None of speech, money, or generosity is a physicalistic concept: judgments in which they're applied are anomalous with respect to the physical. They all involve norms not relevant to the physical

sciences[116]: we might, in a Kantian vein, hold that meaning, money, and generosity are intelligible, rather than sensible, phenomena.

All beliefs about these intelligible, not sensible, phenomena are based, inferentially, on physicalistic beliefs. This is a problematic claim. If the application of these concepts is supported by physicalistic beliefs, then why are these concepts not physicalistic? I discuss this problem at some length in the next section. For the time being, I want to take for granted that concepts of intelligible phenomena can be applied on the basis of physicalistic evidence without being physicalistic concepts.

To support the claim that physicalistic evidence is relevant, I want to discuss how other beliefs holistically support evaluative judgments. We may distinguish between two ways in which supporting beliefs might support. They might be background conditions required for a situation to cause a belief, or they might be the premises of an argument implying the belief to be supported. The difference is that the first kind of support consists in bringing it about that a prompting situation causes a belief; the second kind of support consists in causing a belief. The first kind of support is like the support that oxygen gives the striking of a match, support that allows it to light the match; the second is like the striking of the match.

Consider a simple observation predicate, like '...is red.' We're able to apply that concept pretty directly on the basis of experience. But, as a consequence of the holistic nature of content, believing that something is red does require a host of supporting beliefs. With '...is red,' the supporting beliefs concern the context of observation. A situation might prompt me to believe that such-and-such is red, or not, depending on whether I believe that this is a situation in which normal laws hold. If I believe that I have recently taken LSD, I might discount the prompting of experience on the basis of the fact that I'm in a causally abnormal situation in which situations will prompt me to have beliefs that are false in them.

---

[116] See Davidson 1970, esp. pp. 221-2.

Interpretation is not like that. As I argued in section 2.1.1, interpretation requires supporting beliefs that imply, or at any rate provide logical support for, the interpretation. For instance, the belief that he said "such-and-such is red," plus the appropriate T-sentence for the quoted sentence, imply the interpretation that he said that such-and-such is red. The T-sentence is an empirical generalization embedded within the synthetic *a priori* structure of Tarskian truth theories. The belief that he said "such-and-such is red" is a straightforward empirical belief about sounds coming from a speaker's mouth.

Likewise with moral observations: moral observations call on logical support, not merely the sort of support that would allow a prompting situation to prompt it. The morally weighted situation prompts various non-moral beliefs about it; those beliefs imply or support the moral belief about the situation. Now, I want to say that this claim is just obvious. For what is the alternative? The alternative is to suggest that we can just *see*, directly, something's moral qualities. Wouldn't that require some sort of magic?

Many might cite Wiggins and McDowell as philosophers who believe that we can, in a suitably sophisticated sense of "just see," just see moral properties. McDowell treats moral properties as analogous to secondary properties:

> The idea of value experience involves taking admiration, say, to represent its object as having a property that (although there in the object) is essentially subjective in much the same way as the property that an object is represented as having by an experience of redness — that is, understood adequately only in terms of the appropriate modification of human (or similar) sensibility." (McDowell 1985, p. 143)

"Sensibility," obviously, is broader than perception. If McDowell disagrees with me here, his contention is that moral judgments can be observational judgments, that is, judgments made as a consequence of the direct prompting of moral facts.

Phenomenologically, that is a plausible idea. We do seem struck by moral facts, and we often don't explicitly reason to them. However, often we *do*. Since moral concepts are routinely applied only on the basis of reasoning, it seems more plausible that we always apply them on the basis of reasoning, but that in many cases the reasoning is too

quick and automatic to register phenomenologically. One might reply that we sometimes reason to the application of observation concepts like redness. But the basis of such reasoning is entirely different. When I judge that something is red, if there is evidence for my judgment, it must be evidence from testimony or something of the sort. One doesn't, as a rule, consider other properties of a thing that one currently experiences and deduce from them that it must also be red. But that is just what one does to reason to a moral judgment.

We may say that moral judgments, made however rapidly or automatically, are based on evidence. But not just any set of observations can count as evidence: moral predicates can't be applied to just any objects. In a famous passage, Davidson says that, in interpretation, "...we will try for a theory that finds [our interpretee] consistent, a believer of truths, and a *lover of the good* (all my our own lights, it goes without saying)." (Davidson 1970, p. 222. emphasis added) The same point about the external content of moral judgments that applies elsewhere in interpretation, applies also to the application of moral predicates. Philippa Foot discusses this fact. She wonders whether the application of moral predicates could be externally related to evidence, or whether the relation is internal. If the relation is internal, then we have no understanding of moral concepts independent of the sort of evidence we actually rely on in applying them. If the relation is external, then moral concepts can be applied to unusual objects on the basis of evidence that we wouldn't, ordinarily, recognize as evidence.[117] Can moral predicates be meaningfully applied to just anything? Foot explains:

> On this hypothesis a moral eccentric could be described as commending the clasping of hands as the action of a good man, and we should not have to look for some background to give the supposition sense. That is to say, on this hypothesis the clasping of hands could be commended without any explanation; it could be what those who hold such theories call 'an ultimate moral principle.'
>
> I wish to say that this hypothesis is untenable... (Foot 1958, p. 112)

[117] While the line here is that the relation is internal, the relation is, in truth, only defeasibly or normally internal. The next section will be more precise on this issue.

Foot argues that it's literally nonsensical to ascribe to someone the belief that clasping and unclasping one's hands repeatedly is morally good, without also attributing to him some unusual beliefs about the effects of the clasping. In radical moral interpretation, we would count it against an interpretive scheme if it implied that the speaker believes that repeatedly clasping and unclasping one's hands is good, unless we attributed complementary beliefs about the context (such as that clasping is a conventional sign of respect which is warranted in this context). The belief being attributed to the speaker is nonsense, and we almost certainly misinterpret when we attribute nonsense.

The argument concerns the thin moral predicate '...is good,' but of course it must doubly apply to the thick moral predicates. It's aggressively senseless to apply '...is generous' to an act that has nothing to do with helping someone, or '...is courageous' to an act that isn't done despite fear. Note that I'm not just *stipulating* what moral beliefs we must attribute, or which are true. I'm applying a general theory of meaning and interpretation to the moral case. The argument begins, not with *ad hoc* claims about morality, but with considerations about what a theory of meaning must look like.

Scanlon agrees with the claim that I'm making:

> A reason is a consideration that counts in favor of some judgment-sensitive attitude, and the content of that attitude must provide some guidance in identifying the kinds of considerations that could count in favor of it. If it does not, then the question of whether something is a reason for it will make no sense, and any answer will seem truly arbitrary. (Scanlon 1998, p. 67)

Scanlon's argument is that we would lose any sense for the content of a moral judgment — we would be unable to interpret it — if it lacked the appropriate holistic connections to just the right kind of evidence. The evidence, presumably, will consist largely in normatively relevant, but non-moral, beliefs about the act or object to be evaluated.

What's to have been established is that a principle of charity applies to moral interpretation. To attribute too much moral confusion or falsehood undermines an interpretation. Since I can attribute too much falsehood neither to myself nor to those I interpret, I must find massive agreement between anyone I interpret and me. But it's

commonly held that there is massive, radical moral disagreement; there might even be terminally open, unanswerable moral questions. As David Cooper puts it:

> ...there is a meagre number of significant moral sentences generally held true among us.... If there is this lack of consensus, and if it is taken as a lack of genuine moral consensus, the Davidsonian approach must, it seems, be impaled on one of two horns of a dilemma. Either we stand by the theory of meaning, in terms of agreed applications of terms (of sentences generally held to be true), in which case it looks as if moral terms have lost meaning; or, we do understand moral terms despite the lack of consensus on their application, in which case we must jettison the theory of meaning. (Cooper 1978, p. 105)

Naturally, I reject the dilemma, because I accept the Davidsonian theory of meaning but also claim that we understand moral discourse. The premise of the argument concluding in the dilemma was that there is a lack of consensus on moral issues. I reject the premise.

There is almost universal agreement on almost all evaluative moral judgments that could be made. Unfortunately, most of those judgments aren't ever made; we tend to produce moral evaluations only where there's a controversy. But, while there is moral controversy, it only makes sense against a backdrop of massive agreement. Christine Korsgaard puts the point much as I would. If we encountered moral aliens trying to study us, the fact that they engage in action (like study) implies that we share moral concepts:

> The exact shape of their [moral] problems may be different from ours, and so they may have different conceptions. But they will have views about what is right and what is good, and their language will have terms in which these views are expressed. So we will be able to translate our own terms into their language, and to talk to them about the right and the good. And if we can come at least to see their conceptions as solutions to the normative problems that *they* face, there will even be a kind of convergence. (Korsgaard 1996a, p. 116)

The basic point that Korsgaard is making is that, if someone else engages in action at all, they must have our moral concepts. To engage in action is, as a rule, to do what one thinks best or most right, and one can't think something best or most right without sharing our concepts of the good and the right. Of course, there will be a divergence in application. But the divergence can largely be explained by difference of situation, and

we would agree, on reflection, with most of the relevant divergent applications once we grasped the non-moral facts.

I don't want to just *directly* appeal to difference of non-moral context, or contingent error, to account for moral diversity. What I want to do is take an example of a spectacularly immoral practice, and try to make some sense of it by using our familiar moral concepts. The practice has to be one that's actually defended, on moral grounds, by its practitioners: hypocritical practices won't do. The familiarity of the concepts in terms of which we interpret the alien utterances is the fact that we can interpret them with axioms of Tarskian theories by, more or less, equating the divergent alien's words with words of our own. Thus the predicates must be mainly applied to the same things to which we would apply their translations in our own language. The practice has to be one that we don't think justified by a difference in context between us and the people that engage in it; ideally, it will be their response to a moral question that we also face.

I take the example of honor killings. An honor killing is a killing, typically of a woman, by her family on grounds of sexual dishonor. This barbaric practice is uncommon even in the societies that engage in it, but it's most widespread in the broader Muslim world (though the Qur'an forbids it). Honor killings are done on many grounds. If a woman is killed on grounds of adultery, we can grasp the error. An adulterer or adulteress has violated an agreement, a peculiarly intimate and important one. It's not unreasonable that such a person should be in some way punished. Of course, the punishment shouldn't be death. But we can grasp how the punishment might have been exaggerated beyond reason; we can grasp it with reference to our understanding of the sorts of emotions adultery leads to. Here, we may understand the honor killing as a punishment, because the honor killers' concept of punishment is the same as ours. They've tragically misapplied it.[118]

[118] Those who oppose the death penalty, and yet understand arguments for it and sometimes even sympathize with them, obviously grasp the idea of a tragic misapplication of the concept of punishment.

More bizarre are cases in which unmarried women are killed for having sex out of wedlock. But again, we have a sense for what's going on, with reference to a traditional conception of the commitments of sexual intercourse. Conservative sexual moralists[119] speak of the unity of the human person, and the integration of two persons in intercourse. Unity may be understood in two ways: there is a unity between the mind and the body, and of the person across her history. To marry someone is (usually) to commit to sexual intercourse with him or her. But sexual intercourse integrates two people; it integrates their bodies, and, by way of the unity of the person, it integrates their minds and thus becomes a matter of moral commitment. The two people who are so integrated are themselves unified across time, and so the integration should be, in some sense, permanent. But a premarital integration with another person blocks the totality of the commitment, in a way not entirely dissimilar to adultery. So premarital sex is not entirely unlike cheating on one's spouse, in this conservative view. If adultery warranted punishment, then so should premarital sex. That this argument is very problematic isn't relevant to the point that I'm trying to make, which is only that we can comprehend, in our own moral terms, the thoughts of those alleged aliens that affirm honor killings.

The final, most bizarre, and hardest case is this: a woman might suffer an honor killing *on account of being raped*. That can happen even if those ordering or carrying out the honor killing agree that their victim has been raped. How can one possibly cram this perversity into our moral thinking?

I want to note the problem that would emerge if we were to give up here. We would like to be able to condemn, as immoral, the practice of honor killings. As a practical matter, we would like to be able to convince cultures no longer to engage in it; as a theoretical matter, we'd like to be able to think of them as culpably immoral if they still do. But, if we can't figure out what honor killers are even thinking, then it's not plain what basis we could have for talking with them or condemning them. We couldn't talk

[119] ...such as Vincent Punzo; see Punzo 1969, pp. 192-201.

with them, because we can't understand them; we can't really condemn them, since they don't so much as have the concepts necessary to see the wrongness of what they do. It's crucial that we comprehend these crimes in moral terms, because if we don't, we're reduced to rage or violence in the face of them — and we won't even be able to blame those at whom we're enraged. We would have to treat them as mad, or animals.

How can we comprehend any justification for the idea that capital punishment fits the "crime" of being raped? Let me begin with an assumption. I assume that we can understand the moral philosophy of Immanuel Kant. Pakistani or Turkish honor killers might be beyond our ken, but Kant fits firmly within our western ways of thinking. If we can understand Kant, and Kant has a view that would at least nearly commend or allow honor killings, then we can understand those who commit honor killings. Kant has this to say on the subject of avoiding rape:

> At the moment when I can no longer live my life with honour, and become by such an action unworthy of life, I cannot live at all. It is therefore far better to die with honour and reputation, than to prolong one's life by a few years through a discreditable action. If somebody, for example, can preserve life no longer save by surrendering their person to the will of another, they are bound rather to sacrifice their life, than to dishonour the dignity of humanity in their person, which is what they do by giving themselves up as a thing to the will of someone else. (Kant 1997, p. 150)

It is better, Kant says, to die than be raped.[120] The superiority of the one choice over another is not an experiential, but rather a moral, superiority. Now, this view is preposterous. But we can get a grip on it.[121] There's no question but that Kant's moral language is our own. He misapplies moral concepts that he shares with us.

Kant's view is not identical to that of the honor killers. But we can see how the development would go. It's better to die than be raped; it doesn't matter too much by whose hand one dies. It's better to die than to live with the stain, the indignity, the

---

[120] Though see also Kant 1997, pp. 145-6, on Lucretia. Kant doesn't say that it's better for the victim of rape to commit suicide than not. He doesn't say that it's better to commit suicide than be raped, rather that it's better to be murdered than raped. So he doesn't sanction "honor suicide."

[121] And some people do try to get a grip on it; see Soble 2003, pp. 55-6, and Cooley 2006.

dishonor, of the degradation of one's own humanity. Even more preposterous than Kant's view, but, again, comprehensible in our own moral language.

What has been the point of this example? Cooper offered an objection to Davidsonian moral realism, with reference to the diversity of moral views. Such diversity threatened the idea that we can understand the moral ideas of others. If we can't understand the moral ideas of others, then moral realism has become an implausibly chauvinistic doctrine.

Typically, at this point in the dialectical interplay between realist and relativist, the realist claims that factual and other error can account for moral diversity. That's true. In this case, it's (psychologically, rather than evidentially, motivated) factual beliefs about the superiority of men over women, and over the preferences of the divine, that account for the mistake. But before error can be introduced to account for diversity, it must be established that the concept of error has purchase. It makes no sense to apply the concept of error across conceptual schemes, so what must be shown is that there is no multiplicity of conceptual schemes: that we can understand, in our terms, the discourse of another. That task was basically performed in 3.3, when I gave Davidson's argument against conceptual schemes. However, I wanted to make the point more concretely in the context in which relativism holds most sway.

What is the result of this section? I've suggested that the interpretation of moral predicates, both thick and thin, is sensitive to the objects to which the predicates are applied. If we wish to interpret some predicate as meaning, in its language, what '...is courageous' or '...is right' means in ours, then the speakers of that language must, as a rule, apply their predicates to the same objects to which we apply our corresponding predicates. The diversity of moral views doesn't undermine this claim, because even the most horrific actions, if given moral justifications at all, receive moral justifications that we can comprehend in our own moral terms.

5.2.3 Holistic Aspects of Moral Interpretation 1: Moral Theory

The goal of this section is to determine the role of principles and theory in moral reasoning. I begin by canvassing the end-points of the spectrum.

Jonathan Dancy is the most prominent proponent of particularism, and his view is one of the more extreme ones available. His view is that "the possibility of moral thought and judgment does not depend on the provision of a suitable supply of moral principles." (Dancy 2004, p. 7) A moral principle would be a universally quantified claim connecting satisfaction of some predicates to the satisfaction of thin moral predicates. Why are there no such principles? Dancy appeals to holism about reasons: "...a feature that is a reason in one case may be no reason at all, or an opposite reason, in another." (*ibid*) Let me offer an example. A week ago, I had finished a section of reading, but I needed time to let some new ideas work themselves out. I needed to stop focusing on them. Getting together with friends would distract me from work, and that's exactly what I needed at the moment. So, under those circumstances, that getting together with friends would distract me counted in favor of doing so. Now, though, I'm in the thick of writing and I'm pressed with a deadline. I've had a chance to reflect, and I need to focus on getting this section written. It's very important that I not be distracted. So now, the fact that getting together with friends would distract me from work counts against doing so. Thus there is no principle governing distractions that tells me that its distractingness counts in favor of, or against, doing something distracting. The significance — "valence" — of the distractingness is determined contextually.

Whether to gather with friends is not usually a moral question. In ethical particularism, the idea is that there are no principles connecting any other concepts with the thin moral concepts. No feature of an act counts, in a principled, systematic way, in favor of or against the goodness or rightness of that sort of act.

Considering an obvious challenge offered by Jackson, Pettit, and Smith can sharpen this point:

> [Particularists] might say that all that the right actions have in common is that they belong to the set of right actions. Grasp of the predicate 'is right' simply consists in a grasp of the various [objects] which constitute that set. But this cannot be *all* that unites the class of right actions. There must be some commonality in the sense of a pattern that allows projection from some sufficiently large subset of the [objects] to new members. If there isn't, we finite creatures could not have grasped through a finite learning process (the only sort there is) the predicate 'is right'. (Jackson, Pettit, and Smith 2000, p. 87)

Here is the idea. It's admitted on all sides that the moral supervenes on the non-moral. So if a moral predicate applies to an object, it does so in virtue of a set of non-moral predicates that also apply to the object. Particularism claims that, while that is true, each application of a supervening predicate might well have a different base. The contention is that, if that were the case, then we could never acquire the supervening concept. Since it could result from any of infinitely many subvenient bases, and we would have to learn all of them to grasp the concept, we could never grasp it.

Dancy's reply is to appeal to models of conceptual competence that, he says, don't involve "cottoning on to a pattern that is expressible at the level of the grounds." (Dancy 2004, p. 111) This lack of patterns is the essence of particularism, and, as I'll show, it presents a peculiarly difficult problem for my Davidsonian view. Particularism has it that moral predicates have no systematic, principled relation to non-moral predicates. There is anomalism between the two conceptual realms.

Jackson, Pettit, and Smith contend that, if we can't reduce the moral to the non-moral with principles, then we can never grasp moral concepts. Thus, we can work the reduction. Margaret Little offers the nice reply that we learn, as it were, a reducible fragment of the concept, but then grasp the point of the concept and can leave the reduction behind: "Once we have come to 'catch on' to the concept, though, we are able to discern the very different shape [moral concepts] take in different contexts. To think we cannot is to confuse the conditions of learning with the content of what is learned." (Little 2000, p. 283) While we might have to learn moral concepts in an impoverished, nomic

way, that doesn't imply that the concepts are really nomically connected with the non-moral bases of the application of the concept.

Little, then, is trying to have both anomalism between the moral and the non-moral, and also the acquisition of moral concepts. She continues by offering Davidsonian support for particularism, noting that particularism says nothing about morality that Davidson hadn't said about the mind. Recall my remark about moral judgments, like interpretations and the identification of money, being anomalous with respect to physicalistic beliefs. Little agrees:

> I suspect that holism is a unifying feature of the evaluative. I suspect, that is, that there is no way of cashing out propositionally the ways in which nonevaluative properties contribute to the evaluative natures of situations, actions, characters. (Those persuaded by Davidsonian considerations about interpretation will have theoretical backing to explain this unifying tie: whatever other connections between the evaluative and nonevaluative domains we might acknowledge, there is an essential *anomological relation* between them, where this is read as rejection of articulable laws, because each domain's concepts answer to distinct substantive interpretational constraints.) (*ibid*, pp. 283-4)

Little suggests that holism is a common feature of phenomena governed by norms. But it's not obvious that this holism should everywhere play out as particularism would suggest. It's as a consequence of the holism of the mental that attributing to someone a certain belief counts as evidence that she has some other belief; for instance, if someone believes that grass is green, she probably also believes that grass is a plant, grows on the ground, and so forth. This fact applies across the board, to every speaker in every case.

Particularism will not appeal to holism in remotely the way that Davidsonian interpretation appeals to holism. Davidson appeals to holism to unify and integrate a speaker's utterances and attitudes around a set of norms shared by all. Particularism appeals to holism to eliminate unity and integration of moral judgments in favor of irreducible diversity.

Dancy rejects the claim that this diversity is problematic. Specifically, it does not show that moral judgments have no justificatory or cognitive relations to one another:

> I claim... that one may find other cases instructive or suggestive, whether one is considering simply what features are acting as reasons there is just looking at how things stand there overall. The sceptical challenge is that this [is] just bluster, and that the effect of an extreme holism of reasons is an extreme epistemological atomism. (Dancy 2004, p. 157)

Dancy's epistemological atomism is, we might say, not extreme. Moral judgments can have a wide variety of cognitive and justificatory relations to one another other than relations of implication. By reflection on cases, I can be sensitized to non-moral features that sometimes have normative significance. For instance, by reflection on a case in which tactlessness counted against an act, I can become sensitized to the normative significance of tactfulness in other cases.[122] What I don't do is learn a principle telling me that their tactfulness always counts in favor of tactful actions. In some cases, their tactlessness counts for actions; in some cases tact is morally inert.

One might think that a concept like tactfulness is the concept of an Aristotelian golden mean. The golden mean has a sort of Goldilocks rightness to it: the generous act is the one that is neither too miserly nor too liberal, but *just right*. Shouldn't there be principles connecting the golden mean to the right? No. An action can be evaluated along many possible dimensions, which might be orthogonal to one another. Assume that some action is measurable on the miserly-generous-liberal scale, and also on the cowardly-courageous-foolhardy scale. Being at the golden mean along one such scale doesn't imply being at the golden mean along the other. Perhaps, in this case, generosity would be cowardly and the cowardliness of the action trumps, or even reverses, the generosity. So Dancy could withstand the idea of golden means.

If moral reasoning doesn't involve the application of principles, how do we discover, justify, and explain moral judgments? Rather than knowing principles and how to apply them, what we learn in moral education is a skill:[123]

---

[122] Dancy's example from his 2004, p. 157.
[123] Beyond the passage quoted in the text, see Little 2000, pp. 296-7.

> Particularists conceive of the knowledge brought to a new case as much more like knowledge-how than like knowledge-that. That is, it is a skill of discernment, not knowledge of a set of true general propositions discovered by thinking about previous cases and applied somehow to new ones. (*ibid*, pp. 142-3)

The question whether moral cognition involves the application of principles, or is an inchoate and inarticulable skill, is an hermeneutical question. How shall we interpret moral judges? Shall we take their moral judgments to logically implicate a body of moral theory and principles, or shall we take their moral judgments to flow from an inarticulate ability?

Dancy and Little, it seems to me, make a mistake about cognition's appeal to principles. Dancy speaks of "explicit rule[s]", and says that "One can only compute the articulable." (*ibid*. p. 108) This stress and claim seem to me to be mistaken. There's no problem with appealing to implicit, inarticulable principles. After all, the basis of my view is a concept of truth that, I insist, cannot be articulated. We can't state whatever principles govern our application of the concept of truth. But that's not a reason to treat the application of that concept as being in no way rule-governed. Dancy notes that:[124]

> What it is to go on in the same way need not be capturable in any rule, and what we bring to the new situation need not be an implicit grasp of a suitably context-sensitive rule, but simply an understanding of the sort of difference that can be made by the applicability of this concept, and an ability to apply that understanding to cases that are quite different from the ones in which we originally learnt the concept — an ability that in no way requires either the existence or support of a rule. (*ibid*)

At best, Dancy can show that we *need not* think of moral cognition as involving the application of implicit rules. He can't show that we *must not* think of it in this way. The question is an hermeneutical one: what interpretation of an agent makes the best sense of her actions? Phenomenology will be of no help, since the principles are supposed to be implicit: whatever evidence we appeal to here will need to be either behavioral or flow from the structure of interpretation itself. I offer behavioral evidence below, but here I want to make a point about interpretation.

---

[124] In addition to the passage in the text, see Little 2000, pp. 292-3.

Imagine a robustly skill-governed situation, a basketball game. We see player A duck left, feint right, and then pass the ball to her teammate C, and then we try to offer an interpretation of her play. Imagine how unsatisfying it would be to offer, as an interpretation, only "She knows how to play basketball," or "She's a skillful player." Surely we can do better. But what would a satisfying interpretation look like? It would appeal to various beliefs of the player at different levels of generalization. At the lowest level are beliefs like that she has already dribbled, so she must pass or shoot; that she's such-and-such a distance from the basket and covered by opponent D; that teammate C is over *there* and uncovered while teammate B, in whose direction she had initially ducked for a pass is over *here* and covered. But a satisfying interpretation will then appeal to general beliefs. She doesn't shoot, because she knows that, as a rule, she couldn't make the shot from here and that being covered by an opponent (such as D) lowers the odds even more. She doesn't pass to B, because B is at a position from which a shot is difficult and besides, the opponent covering B, as a rule, could intercept from that sort of covering position. She passes to C, because, as a rule, it's better to pass to uncovered teammates; also, C is at a position advantageous for a shot. To interpret the employment of the skill, we must attribute a set of beliefs. To offer the skill itself as the last word in interpretation is to offer a fairly unsatisfying interpretation.

Here's another way of putting the point. Particularism is threatened by extreme epistemological atomism: the problem that no two cases will have any bearing on one another. The response is to appeal to non-logical connections between cases, such as suggestion or inspiration. Taking the right sorts of suggestions and inspirations will constitute a skill. Particularism thus replaces principles with skills as the unifying factors of moral reasoning. The great advantage of principles over skills is that we can gather the sense of someone's actions if we attribute belief in principles, but not if we attribute inchoate skills. The inarticulability of skills turns them into explanatory and

hermeneutical black boxes, preventing us from grasping the state of mind of those whose actions we would like to interpret.

When we must offer a skill as explanation, we do in fact surrender intelligibility. The player's choice of whether to shoot or pass, and to whom to pass, was one that we could comprehend. However, the way the player moved her fingers when she passed probably overreaches what we can account for in specifically rational terms. The pattern according to which the palm and fingers begin concave but rounded around the ball, and end concave but more conical, having pushed the ball through space, is not a pattern that we could rationalize. If moral discernment is unintelligible like that, then Dancy is right, but I can't help but think that our moral discourse would be the poorer for it: poorer than it is.

Could my suggestion that there are inarticulable principles be any better? Yes, it can. The principles to which we appeal in moral reasoning are not fully articulable. But, I will suggest below, there are defeasible, contextual specifications of them that are articulable. Most of our moral reasoning appeals to these articulate principles.

The opposite extreme is generalism, and the generalist view that I want to address is a Kantian one. For Kant, application of moral principle is not a mere discovery procedure in ethics: it's constitutive of the ethical. Only in adherence to rules can we find autonomy, and only in autonomy can we act morally, as distinct from according to non-moral inclinations, however prudent.

As Rawls explains, Kant's view is centered in the ideal of autonomy, which is to regulate all practical life:[125]

> ...the order of moral and political values must be made, or itself constituted, by the principles and conceptions of practical reason. Let us refer to this as constitutive autonomy. In contrast with rational intuitionism, constitutive autonomy says that the so-called independent order of values does not constitute

[125] See also Rawls 2000, pp. 226-37; also, of course, Kant 1785/1949, esp. pp. 56, 63-80. (orig. pagination, pp. 71-2, 78-99.)

> itself but is constituted by the activity, actual or ideal, of practical (human) reason itself. (Rawls 1993/1996, p. 99)

What would make such an image of the ethical attractive? Kant's ethics engages with us as we face choice.[126] Construed in physicalistic terms, actions of big bundles of particles like us are physically determined. But the fact that my actions are physically determined doesn't alleviate the problem of choice. I face choice not as a bundle of particles, but as a rational agent. Agency involves rationality, which is constitutively normative. So I face choice not as a physical object, but only insofar as I satisfy predicates that do not appear in, and cannot be reduced to, the language of physics. But, since it's my agency in virtue of which I face choice, it makes no sense to look to facts independent of my agency to determine what to do. To look to physical facts or pre-rational urges to determine my acts is to pretend that I don't actually face a choice. Insofar as I come to grips with the fact of choice, I have nowhere to look for advice but myself. If my decisions were random and unpatterned, I could hardly be said to be choosing. If I aim to understand what I do, then I have to make myself interpretable: my decisions must follow according to some sort of rule that I can understand. My actions have to cohere as a whole. Since the only guidance I have is that I must act according to principles, the only principle I can follow is to act according to something that can *be* a principle: a rule that I could follow, and have others follow, consistently. Following such rules is constitutive of rational agency and moral behavior.

I find this image attractive and persuasive — overwhelmingly so — despite the clumsiness of my expression of it. But if principles are constitutive of moral reasoning, how can we accommodate Dancy's plausible claim that a feature of an action can count in favor of it on one occasion while, on another occasion, the same feature might count against some other action that would also have it? If moral reasoning is constituted by principles, there should be universally quantified connections between the non-moral

[126] See Korsgaard 1996, pp. 94-108.

features of a possible act and its moral features. What else could a principle be? What else could I attribute to someone in order to have a rational interpretation of her acts?

Davidson's Kantian position about the attitudes is that they are constitutively normative. However, the brain states that subvene those attitudes are not constitutively normative. There is thus anomalism between the mental and the physical.[127] Particularism (especially in Little's interpretation) has it that the moral is also constitutively normative, unlike its subvenient base. So far, one would like to just accept both views. However, the holism of the mental has it that the constitutive normativity of the mental is its principled nature. Thus moral beliefs can't be identified unless they adhere to principle. Identical considerations drive both anomalous monism and particularism, but anomalous monism is in tension (at least) with particularism.

I want to consider a related question of particularism about meaning. Dancy is a holist about linguistic meaning: a word's contribution to a sentence is determined by the context of the sentence. While Dancy does accept "weak compositionality," the claim[128] that sentence meaning is determined by word meaning, he does not accept "strong compositionality," the claim that a word contributes the same meaning to each sentence in which it appears. Rather, each contribution may be different.

Jackson, Pettit, and Smith would reject Dancy's claim. There must be a finite number of possible contributions that a word could make; else, the word's meaning would be beyond our power to grasp. Dancy again appeals, albeit tacitly, to skill, this time in the guise of competence:

> ...would not this mean that the term is essentially ambiguous? It looks as if terms, on this account, change their meaning as they move from context to context, and this is surely hard to swallow. To answer this, we need to remember that the meaning of the term is what one knows when one is a competent user of that term. If the term is capable of making a range of contributions in differing contexts, this is part of what the competent user must know.... The meaning of the term,

[127] See Davidson 1970.
[128] Dancy 2004, p. 193.

> understood in general, is the range of differences that it can make... (Dancy, 2004, p. 194)

Here is a point that I can grapple with directly on the basis of the linguistic and hermeneutical considerations of earlier chapters. Dancy's claim is this. A term has one, and only one, meaning. But that meaning is all of the possible contributions that it could make. In one context, a word is such that it makes a certain contribution, but, in that context, it's also such that, in other contexts, it would make a different one. So the fact that a word makes a different contribution to each sentence in which it appears doesn't show that the word is ambiguous.

From an interpretive standpoint, the last component of the claim won't do. It undermines our grasp of real ambiguity. A term's ambiguity is just the fact that it makes a different contribution in different contexts; changing the traditional 'meaning' into 'contribution' only required that we rephrase the traditional notion of ambiguity in the fancy new jargon. And then it turns out that Dancy has it that terms are systematically ambiguous. But they aren't. Obviously we sometimes use ambiguous language, but just as obviously, usually we don't. Note that Dancy has to know that words make systematically related contributions in different contexts; else, he would just urge that terms *are* systematically ambiguous.

Imagine that we try to interpret some speaker, and there appears to be no pattern at all to her use of a term: it seems to make a different contribution to every sentence in which it appears. We would begin to wonder if we were missing something. Perhaps it isn't the same word in each context, because the language we're trying to interpret has tonal qualities that we hadn't noticed. Or something. We would lose our sense of the word being the same word from case to case if its use weren't patterned.

Note that this isn't just a fact about language learning. It's a fact about interpretation in general. Consider that Dancy's sober assessment of language is *more extreme* than what one might say to *satirize* a language as effectively unlearnable:

> There are some exceedingly useful words in this language [German]. *Schlag*, for example; and *Zug*.... The word *Schlag* means Blow, Stroke, Dash, Hit, Shock, Clap, Slap, Time, Bar, Coin, Stamp, Kind, Sort, Manner, Way, Apoplexy, Wood-cutting Inclosure, Field, Forest-clearing. This is its simple and *exact* meaning — that is to say, its restricted, its fettered meaning; but there are ways by which you can set it free, so that it can soar away, as on the wings of the morning, and never be at rest. (Twain, 1880, p. 1150)

The trouble with *Schlag* is that it isn't fettered by principles. Its contributions to the sentences in which it appears are too diverse for the word to be learnable. Notably, Twain is making a joke. Of course, German isn't really as awful as he says. No language could be.

Dancy says that the meaning of a word is its range of possible contributions. That's true. The question is whether that range is small enough to be captured in a disjunctive Tarskian axiom, for such axioms are the principles that we would apply in linguistic comprehension and use if we apply principles at all. The axiom can even be rather long, consistent with the view that principles govern linguistic comprehension. If not, linguistic comprehension and use is a rather depressing and inchoate skill. I want to directly apply Davidson's theory of meaning here. Dancy's suggestion about language conflicts with the best theory of meaning on the table, so it's probably wrong.

Of course, a word can be pressed into service for a wider variety of uses than the ones in its Tarskian axiom, however long the axiom. Doesn't that confirm Dancy's view? Not precisely. Dancy's claim is that we don't need principles to govern meaning. But we do. For in the absence of a relatively stable core of possible contributions, we couldn't identify a word as meaningful at all. Thus we wouldn't be able to press it into service for unusual purposes, such as puns, jokes, and other parasitic and aberrant uses. If Dancy opted for an extreme, and permitted any word to have any meaning in an appropriate context, we would ask about parasitic uses: Why choose *that* word, when any word can have any meaning? Dancy's view is not so extreme, but also not so moderate that a more moderate form of the question wouldn't be problematic for him. Dancy would be right if

he said that no Tarskian axiom can capture *every* use of a word. But he's wrong to say that principles are entirely unnecessary to linguistic comprehension and use. This argument against particularism about language presages my argument against particularism in general.

I will propose that moral exceptions, situations in which our usual principles would lead us astray, are necessarily statistically abnormal, in much the same way that hallucinations and incoherent beliefs are statistically abnormal. Most morally relevant features have a standard valence, which shifts or becomes irrelevant only in fairly unusual circumstances. Dancy is sceptical of the idea of default reasons, that is, considerations with a standard valence that applies in all normal cases.[129] He accepts that such reasons could possibly exist, he just doesn't know of an argument showing that they do. He is specifically sceptical of my view:

> The principle that it is wrong to lie cannot be merely a generalization, a claim that lies are mostly the worse for being lies, for if all moral principles were of this sort, the argument that moral thought and judgment depend on the possibility of moral principles would simply be the argument that such thought is impossible unless there is a preponderance of normal cases over abnormal ones. I have never seen this argument made... (Dancy 2000, p. 76)

That is the argument I offer.

I want to begin with a piece of behavioral evidence for the importance of principles. We do, in fact, appeal to principles, specifically, principles connecting thick moral concepts to thin. McNaughton and Rawling argue:

> That an act is cruel, mean, or dishonest counts against it; that it is kind, generous, or honest counts in its favour. On thin intuitionism [particularism], thick moral properties have no more intrinsic moral significance than non-moral properties. It will, presumably, turn out that these properties are 'commonly more important' than some others (though thin intuitionism owes us an account of why), but that not only understates their force, it seems to mislocate their centrality. It is not just that it is helpful to look at them because they often count; their counting is central to their being thick *moral* concepts. (McNaughton and Rawling 2000, p. 273)

[129] Dancy 2004, pp. 111-7.

McNaughton and Rawling offer the analogy of a "fit" between various concepts. The thick and thin moral concepts fit together as a consequence of the holism of the attitudes.

Consider an attempt to interpret a speaker. Imagine that the speaker seems to believe, each time that a possible act would have feature G and no other morally relevant feature, that there is nothing to be said for the act at all. G cannot be generosity. Its generosity, by its nature, counts in favor of the generous act. That an act would be generous doesn't close the question whether to perform it: maybe I can't afford generosity right now, or maybe my potential beneficiary doesn't deserve my help. Moreover, there can be cases in which its generosity counts against an act. Perhaps someone needs to be left to fend for herself; her problem is that people have been too generous with her, and now she needs some cold reality. In that sort of case, generosity counts against. But usually, generosity is not enabling behavior. Their normal valences with respect to the thin moral concepts are constitutive of the thick moral concepts. Thus there are defeasible principles connecting the thick to the thin.

If the particularist wants to accept that its generosity typically counts in favor of a generous act, then she must explain why. On particularism, there are no principled connections between the thin moral concepts and anything else. But we can see that we won't attribute beliefs involving the thick moral concepts to someone unless we're willing to attribute to her principles connecting those concepts to the thin moral concepts. That the principles are defeasible isn't relevant: they apply in most cases and all normal cases, and someone without the principles doesn't grasp the concepts.

To attribute the belief that an act is generous, then, I must ordinarily also attribute the belief that there's at least something — its generosity — that counts in favor of that act. But I must also attribute other, non-moral beliefs. Assume that the speaker often tries to treat people G-ishly. However, the people she wishes to treat G-ishly are all enormously wealthy, much more so than the speaker herself. Also, they have no special

relationship to the speaker. Treating someone G-ishly is obviously not treating her generously, since you can't be generous to wealthy strangers.

Here is my argument for principles. In interpretation, I must attribute beliefs holistically. There are defeasible but massive inferential connections between non-moral beliefs, thick evaluations, and thin evaluations. These connections are moral principles. If I interpret holistically, then I must interpret as if the speaker respects these principles: if she doesn't, I begin to lose any sense for what she means. So the speaker must *have* principles, since there isn't any difference between having and respecting principles. To be interpreted as one who has moral beliefs, one must be interpreted as one who has moral principles. Contrary to Dancy's particularism, the possibility of moral thought and judgment depends on the provision of a suitable supply of moral principles.

Consider Little's remark that there is an analogy (at least) between Davidson's anomalism about the mental and particularism's anomalism about the moral. Jackson, Pettit, and Smith seem to require that moral distinctions be drawable in non-moral language; that is, their argument for principles is just that anomalism undercuts learnability. My view is intended to save the anomalism of the mental and the moral while simultaneously preserving principles. If the principles were reductions of the mental or the moral to something else, then anomalism would be lost. But the principles are not reductions. They are defeasible. They introduce sufficient regularity in the connections between the mental and the moral, on the one hand, and the non-mental and non-moral, on the other, to make these realms graspable on the basis of behavioral evidence, but not so much regularity that they're reduced to the physical.

Note that when I speak of making the moral realm graspable, I don't mean to be speaking, as Jackson, Pettit, and Smith do, of the *acquisition* of moral concepts. I doubt that we acquire thin moral concepts; I suspect that they're part of an *a priori* theory of agency with which we approach the interpretation of agents. To grasp the realm of the moral here just means to be able to *attribute* attitudes involving moral concepts to

speakers and agents. The moral is anomalous with respect to the non-moral, but there still must be sufficiently strong connection between the two, and among beliefs within the former, for us to be able to treat moral believers' moral attitudes as holistically connected to their non-moral attitudes. Defeasible connectives provide exactly the right strength of connection.

My view is closer to Holton's "principled particularism" than it is to other particularisms on the market. For Holton, different moral principles apply to situations depending on the situations' moral complexity. For instance, if the only morally relevant feature of an action is that it would be generous, then you should do it; if, on the other hand, it is generous but also enabling behavior, and *those* are all of the morally relevant features, then you shouldn't do it.... Each principle is a material conditional with a finite conjunction of morally relevant features, plus a "That's it" clause stating that there are no more morally relevant features, on the left, and a moral evaluation on the right. For instance, such a principle might run: ∀x(x is dangerous and x is courageous and That's it → x is right). Moral reasoning involves concluding with the thin moral evaluation on the right, which follows from the assertion that the morally relevant features in the conjunction obtain, and "That's it," and the principle.[130]

This approach has certain logical oddnesses.[131] For instance, the 'that's in each such argument, even ones that apply the same principle, have different referents. Consider this argument:

(1) ∀x(x is generous and That's it → x is right)
(2) a is generous
(3) That's it
Thus, (4) a is right

Imagine another argument, identical but that its (2') says that b is generous, and its (4') says that b is right. In the stated argument, the 'that' in (3) refers to (2). In the other

[130] See Holton 2002, pp. 199-201.
[131] Holton notes this oddness on p. 201; see also 204-6. I'm not *completely* sure that Holton's self-criticism is correct; if it isn't, then there may be less to decide between his view and mine.

argument, the 'that' in (3') refers to (2'). For (3) to hook up with (1), and (3') to hook up with (1'), the 'that's in (1) and (1') must similarly vary their referents. But, according to Holton, that gives them different meanings; they turn out to be different principles. That is not an acceptable consequence. The point of principles was to give unity to the mental, a unity without which the mental dissolves. If a different principle is called on for each instance of moral reasoning, the principle doesn't appear to unify. Some higher-level propensity for generating contextually appropriate principles would have to do the unifying work. So I think that defeasible conditionals are a superior model.[132]

Recall that, while interpretation is holistic, there is also a structure in which application of predicates directly to objects has greater weight in determining predicate meaning than more abstract, farther-from-experience beliefs. Indeterminacy, likewise, increases with distance from experience. The same structure applies in moral reasoning. Evaluations have greater weight in determining the content of the moral concepts that appear in them than do principles. Without principles, moral evaluations wouldn't be what they are, but the mass of evaluations has greater evidentiary weight in decisions about what beliefs to attribute than principles do. It's easier to determine someone's evaluations than her principles, and the evidence flows accordingly.

There is a further, related point. Defeasible principles don't apply everywhere. If moral reasoning consists in the application of principles to cases, then how do we do moral reasoning where our defeasible principles don't apply? It's true that we can't, in these cases, apply the principles. However, the morally relevant predicates that are present must also appear in inapplicable defeasible principles. In grasping those principles, we see the point of these predicates. That gives us an insight, though one that isn't rationally graspable or interpretable, in the exceptional cases. Ideally, it would give us an insight that we could then codify and justify, thus increasing the applicability of our principles and broadening the range in which our moral thinking is interpretable — is

[132] Little appears to offer a similar idea at Little 2000, pp. 299-300.

thinking properly speaking. I would suggest that, in this way, we articulate more and more parts of the non-defeasible principles actually governing our uses and actions. These principles are too complicated to ever be fully articulated — like the concept of truth — but we can articulate their application in particular contexts, just as we can articulate our application of the concept of truth to particular speakers by forming Tarskian theories for them. Where so-far articulate principles don't apply, though, particularist descriptions of moral insight will have to do.

With defeasible principles in place, I can move on to an issue in moral epistemology. The nature of the principles connecting between non-moral facts, thick evaluations, and thin evaluations suggests that reflective equilibrium has a justificatory role to play in moral reasoning.

Particularism has little use for reflective equilibrium.[133] For Kantian generalism, the role of reflective equilibrium is chiefly as a check on speculation. The Kantian believes that an act's rightness is a consequence of its being prescribed by a principle. Principles, then, do the real work. But a modest Kantian might reflect on the dangers of moral theories. How could one check that a moral theory, a body of principles, actually deserves rational acclaim? On the assumption that the rational moral principles are embedded in our actual moral evaluations, the Kantian might conclude that, while evaluations are fallible, they can provide some insight into the validity of principles. Thus we should check our principles against our evaluations in an effort to secure the principles most deserving of rational assent.

In chapter 4, I opposed the sort of expressivist and prescriptivist approaches to moral language that could make it reasonable to secure assent to moral utterances without thinking that they're true. In chapter 1, I tried to show that truth involves a connection to independent truth-conditions. So I disagree with the traditional Kantian generalist that a moral utterance's deserving rational assent consists in its derivation from principles. On

[133] See Dancy 2004, pp. 153-4.

the contrary, we should give our assent to what seems to be true to independent truth-conditions.

As a consequence of the anti-sceptical argument of section 3.2, we can see that our moral beliefs are mainly true unless there is some special sceptical problem for moral cognition. Assuming that there isn't, what more could we ask for in a method for moral reflection than that we make our mainly true beliefs cohere with one another? The broadly coherentist epistemology implicit in Davidson's approach is adequate in morality, too. The fact that moral content tends to flow from evaluations to principles, but that in the absence of principles there would be no evaluations, shows that we need both sort of moral belief. The best way to get coherence, and hence truth, is to check evaluations against the systematizing principles, and contrariwise, until the system makes as much sense as we can make it.

Susan Hurley distinguishes between "centralist" and "non-centralist" accounts of ethical concepts. Centralism has it that the normative content of thick ethical concepts is determined by their relation to the thin ethical concepts; non-centralism disagrees.[134] My view is non-centralist. A thick moral concept is the concept it is because of its principled relations to various non-moral concepts, and to the thin moral concepts; but the thin moral concepts are the concepts they are because of their principled connections to the thick moral concepts. I note this because of a point that Hurley makes about a consequence of non-centralism:

> Non-centralism claims that there are logical connections between claims about what ought to be done, all things considered, and a list of specific values; the sense of *ought* that is a function of the specific values on the list can be used to challenge and revise views about the relationships among those values, but it cannot be used to endorse an entirely new list. Thus, non-centralism threatens to deprive us of a sense in which to disagree about things we seem to want to disagree about. In what sense can we disagree with someone who does not share our specific reason-giving practices? (Hurley 1985, p. 67)

[134] Hurley 1985, p. 56.

Where Hurley speaks of threats, I would speak of promises. The promise of the holistic, coherentist approach is that it puts our moral discourse in relationship with any and all moral discourse. Far from preventing meaningful disagreement, it lays the meta-ethical groundwork for moral engagement. By engaging with foreign moral theories, I come to a better understanding of the moral concepts embedded in both the foreign theory and mine. If I could reach wide reflective equilibrium, I would have taken into account all other moral theories, but our common moral concepts would only have been sharpened and revised, never done away with. Hurley's mistake is to think, as Fodor and LePore seem to,[135] that holism commits us to either total agreement or else incommensurability.

Notably, the argument Hurley offers accounts for the justification one gets from a coherentist approach in moral epistemology. There's no moral justification that won't justify most of my moral beliefs, since what it is to be a true moral belief is, approximately, to be, or to cohere with, my moral beliefs. So I know that my beliefs can't be massively wrong. Thus rendering them coherent is justificatory.

How do things stand with Kantian autonomy? I want to suggest that my view is Kantian: rational agency is autonomous in its moral reasoning, despite its appeal to independent moral facts. For Kant, autonomy was a formal matter, because, for Kant, there was no content that something had to have to count as a mind. The mental was characterized with reference to a structure of categories and schemata. On this approach, the mind has only formalities to fall back on when it looks to itself for rational governance.

Davidson's approach is squarely Kantian, in that it supports the anomalism of the mental. The mental is an autonomous conceptual realm, logically connected to the physical only by defeasible conditionals that don't try to articulate the nature of the mind in physical terms, but only to tie attitudes and contents to the behavioral evidence without which they couldn't exist. Nevertheless, Davidson's approach, in rejecting conceptual

[135] See the discussion back in 2.3.

schemes and in requiring massive agreement for interpretation, gives substance to the mind. A mind necessarily has certain content; anything without that content couldn't be interpreted and wouldn't be a mind. This is a rational requirement, and is itself anomalous with respect to the physical.

If Kantian autonomy consists in a mind not looking beyond itself for reasons to act, then my approach is not Kantian. But I doubt that the core of Kant's approach is this negative idea. I suspect that the core of Kant's approach is the idea that the mind should not look beyond what it's rationally required to accept. Kant thought that all that we *must* accept, on narrowly rational grounds, were certain formal beliefs, analytic or synthetic *a priori*. The Davidsonian approach has a place for the synthetic *a priori*, as I've suggested. But the Davidsonian conception of rationality is substantive: to be rational, a mind must adopt a wide swath of beliefs. That set of beliefs is rationally required of us. So if Kantian autonomy consists in looking to the rationally required, then one can look to one's entire body of well-justified moral judgments and principles for guidance in action.

5.2.4 Holistic Aspects 2: Moral Motivation

In this section, I confront the problem of moral motivation by way of discussing holistic constraints on directive judgments (which, to recall, are specific beliefs about what the thinker herself ought to do in some particular situation). Here is the problem. As I argued in chapter 4, moral discourse is cognitive in nature: moral utterances present beliefs, not desires. However, moral utterances seem to have a tie to motivation. As a rule, when someone believes that she ought to do something, she experiences some pressure in favor of doing it. Nevertheless, belief is typically held to be motivationally inert. So, while the attitudes presented by prescriptive judgments are beliefs, which are motivationally inert, they seem correlated with motivation. What accounts for this correlation?

In general, the problem is one about the nature of reasons for action. To begin, I need to distinguish between two completely different kinds of phenomena that are both

reasonably called reasons for action: motivating and explanatory reasons, or first-person and third-person reasons.[136] A motivating reason is an independent factor that an agent considers when engaging in practical reasoning. An explanatory reason is a set of attitudes the agent has that cause her to act. The motivating reason for performing A is whatever the agent considered to count in favor of doing A. The agent might believe that there are motivating reasons that don't actually exist, as when the agent believes that the possible action would have a feature that it wouldn't have, and counted its having that feature in favor of performing the action. The explanatory reason is the cluster of beliefs, desires, intentions and so forth that cause her to do A. When we reason about what to do, we don't, as a rule, take into account our beliefs and desires: we take into account external facts. But when we try to account for why people do what they do, we don't, as a rule, take into account external facts: we take into account her beliefs and desires.

From the first-person point of view, when considering possible actions, we mainly consider features of that action and its leading competitors. Being desired by us is not, as a rule, a relevant feature. If I decide to get some ice cream, I don't consider my desire for ice cream: I consider the flavor and texture and expense. Now, sometimes we do consider the desire as relevant. As with itches, desires sometimes attract so much attention that we would rather place elsewhere that we concede and satisfy the desire so as to set it aside and return to what we think best.

From the third-person point of view, when explaining actions, we don't consider external facts. Sometimes, people act on the basis of false beliefs and expectations. I can hardly give a causal explanation of someone's action with reference to non-obtaining facts about the action. Since I should appeal to the same sort of factors in explaining

[136] The distinction is drawn, with varying degrees of distinctness, at Nagel 1970, p. 15; Bond 1983, pp. 22-3, 28; Darwall 1983, p. 28; Audi 1986, p. 76; Brink 1989, p. 39; and Smith 1994b, pp. 94-5.

action based on true belief and action based on false belief, I should appeal to attitudes of the agent in both cases.[137]

The question is about the nature of explanatory or third-person reasons, and the main contenders are Hume's theory and Kant's, or, at any rate, theories attached to these thinkers' names. The Humean theory, in brief, has two components. First, every explanatory reason has at least one desire as a component.[138] Second, desires come in two kinds. First, there are basic desires. These desires are psychological contingencies that don't rationally derive from any other attitudes by any form of reasoning. Second, there are derived desires. These desires derive from basic desires. The derived desire is to perform some action that either will help *cause* the realization of a more basic desire, or will itself at least partly *constitute* the realization of a more basic desire. No desire is rationally derived from any attitude other than other desires.[139] Of course, belief plays a role in the reasoning from basic to derived desire, and the derivation of derived from basic desires can be counted as a form of *reasoning*, but the beliefs and rational processes are themselves impotent to do anything but guide desire.

Kant's theory has it that pure reason can motivate. Unfortunately, Kant is none too clear on what's going on. Rawls has it that, for Kant: "...our consciousness of the moral law as supremely authoritative for us must be so deeply rooted in our person as reasonable and rational that this law by itself, when fully known and understood, can be a sufficient motive for us to act from it, whatever our natural desires." (Rawls 2000, p. 255) Kant himself says in the second *Critique* that "What is essential in the moral worth of actions is that the moral law should directly determine the will." (Kant 1788, p. 75; orig. p. 71) How could that happen? On the one hand, Kant says that "...how a law in itself can

---

[137] It would lead me very far astray to explain why there is no analogy between this argument and (bad) arguments from illusion and hallucination for perceptual representationalism. Suffice it to say that there are many disanalogies, having to do with direction of fit, direction of cause, and the sort of content the different states have.

[138] Smith offers this idea at Smith 1994b, pp. 92-3.

[139] Williams offers this idea at Williams 1980, esp. p. 109.

be the direct motive of the will is an insoluble problem for the human reason." (*ibid*, p. 75; orig. p. 72) Now, "a law in itself" can't causally account for action. What Kant means is that *belief* in the law accounts for moral action. And he admits that there's no way to understand how that could happen. But on the other hand, Kant had said a few years earlier in the *Groundwork* that for beings that, like us, aren't entirely denizens of the noumenal realm to "will what reason alone direct such beings that they ought to will, it is no doubt requisite that reason should have a power *to infuse with a feeling of pleasure* or satisfaction in the fulfilment of duty..." (Kant 1785, p. 77; orig. p. 96) To whatever degree we associate pleasure with the satisfaction of desire, this seems to be a concession to Hume. We may set aside such concessions in the name of having a distinct doctrine to call Kantian. For Kant, then, what's central to moral motivation is what Korsgaard calls the internalism requirement: "Practical-reason claims, if they are really to present us with reasons for action, must be capable of motivating rational persons." (Korsgaard 1986, p. 317) I take Korsgaard to mean that moral beliefs must be capable of causing action.

The first question, then, is whether the explanatory reason for an action must, in any interesting sense, include a desire. There is substantial scepticism about this claim. Nagel has it that, while every action is caused by a desire, this fact is a triviality with no explanatory power:

> The assumption that a motivating desire underlies every intentional act depends, I believe, on a confusion between two sorts of desires, motivated and unmotivated. ...many desires... are *arrived at* by decision and after deliberation.... Rational or motivational explanation is just as much in order for that desire as for the action itself.
>
> The claim that a desire underlies every act is true only if desires are taken to include motivated as well as unmotivated desires, and it is true only in the sense that *whatever* may be the motivation for someone's intentional pursuit of a goal, it becomes in virtue of his pursuit *ipso facto* appropriate to ascribe to him a desire for that goal. But if the desire is a motivated one, the explanation of it will be the same as the explanation of his pursuit, and it is by no means obvious that a desire must enter into this further explanation. (Nagel 1970, p. 29)

Bond[140] calls these motivated desires "logical ghosts." Here is the idea. It could be that, in order to account for action, a desire to perform it must be posited. But we may then ask what motivated the agent to perform the action, and also what led the agent to want to perform the action. We will typically offer exactly the same answer to both questions. It's not at all obvious why that answer must itself appeal to further desires. In fact, it will be fairly rare that the agent's considerations, her motivating reason, included any reference to her desires at all. Nagel's line, then, is to accept the first Humean claim as a triviality, and reject the second.

McDowell offers a similar argument:

> Charitable behaviour aims at an end, namely the good of others. It does not follow that a full specification of the agent's reasons for a charitable act would need to add a desire to his conception of the circumstances in which he acted.... The desire for the good of others is related to charity as the desire for one's own future happiness is related to prudence; not, then, as a needed *extra* ingredient in formulations of reasons for acting. Rather, the desire is ascribed... in recognition of the fact that a charitable person's special way of conceiving situations by itself casts a favourable light on charitable actions. Of course a desire ascribed in this purely consequential way is not independently intelligible. (McDowell 1978, p. 84)

McDowell urges roughly the same consideration that Nagel does. When an agent focuses on considerations that favor her performing some action, she might be led to perform the action. If we wish, we may say that she is also led to desire to perform the action, in some formal sense of 'desire.' But the desire we posit is not an autonomous mental entity, and it lacks any real causal or explanatory power. It is a ghost of the actual explanatory factors, which are themselves beliefs that the action has features that recommend it more highly than its competitors.[141]

If we should agree with the first Humean claim, taken as a triviality, what should we think of the second, the claim that all desires are either basic or derived from other

---

[140] See Bond 1983, pp. 9-13.

[141] Bittner, at Bittner 2001 pp. 18-22, offers another argument for the claim that the desires, and also the beliefs, in belief-desire reasons are not independently intelligible and have no apparent independent causal roles.

desires? Nagel and McDowell suggest the Kantian view that belief, by itself, can explain action, albeit by way of causing ghost desires. Are they right? Ross offers this argument for the second Humean claim:

> ...it's not plausible that we can explain the directiveness of desires in terms of relations among intentional states which are not themselves directive states.... The basic point is that it's implausible that we can explain a state's having the causal role characteristic of directive fit, which involves behaving in ways to change the world in order to satisfy a certain representation of it, in terms of states with the causal role characteristic of descriptive fit, which doesn't involve behaving in these ways. (Ross 2002, p. 202)

Ross distinguishes between beliefs and desires according to their directions of fit.[142] Beliefs characteristically try to match the world, and if a belief fails to match the world, the belief is what has gone wrong. Desires try to make the world match them, and if the world fails to match a desire, the world is what has gone wrong (from that desire's point of view, at any rate). If we want to admit that desires are required for action, then we must admit that one of the required states for action is something that seeks, constitutively, to change the world. That is perfectly natural. When we ask why an agent has a desire, we have two choices. We could either posit a non-rational explanation, from physiology or training or what-have-you, or a rational explanation. Desires with non-rational explanations are basic desires; those with rational explanations are derived desires. If we then ask why an agent has some derived desire, Ross says, we must appeal to another desire. A desire, with its characteristic direction of fit, is necessary to rationalize any derived desire. The reason is that it's hard to see how beliefs alone could *rationalize* an attitude that isn't true or false. Beliefs rationalize by implication, and a desire can never be a logical implication, since implication is truth-preserving and desires aren't truth-evaluable. A desire can be rationalized by another when the desirer believes that the new desire is for something that would partly realize or help cause the satisfaction of the old desire; no logical relationships between the desires themselves are

[142] While Elizabeth Anscombe seems to have originated the idea of direction of fit in her book *Intention*, my understanding draws mainly from Searle 1983.

necessary, but a believed relationship of realization or causation between their *contents*, is.

Consider an example from Stephen Darwall[143], offered as part of an attack on the second Humean claim. We are to imagine that Roberta (the Buddha?) has grown up with little awareness of the pain in the world, but that, on maturity, she is shown a film depicting some among the many who suffer. She acquires the desire to ease their suffering. How shall we interpret the acquisition of this desire?

The Humean has to interpret it as the derivation of a derived desire from basic desires, given new beliefs. But it's not obvious to which desire(s) there should be appeal. Derived desires are desires to perform actions that either cause or constitute the satisfaction of the relevant basic desire(s). Relieving this particular suffering would cause the world to be a better place, or constitute the partial satisfaction a more basic desire to relieve suffering in general. Does Roberta have a basic desire for the good? I return to this possibility. Does she have a basic desire to relieve suffering in general? Roberta has never before acted so as to relieve suffering, or spoken in favor of doing so. The only evidence in favor of her having *that* older basic desire is her new, allegedly derived, desire, but whether an old desire is necessary to account for it is what's in question.

Ross just repeats his claim that it's implausible that the new desire could be accounted for without reference to some old desire. I can imagine the more serious Humean taking any of several directions. She might posit a whole set of innate moral desires. Hume, of course, couldn't do that. It's quite impossible to desire that Fa without having the concept of F-ness, and for Hume, no concept is innate. So the concept of F-ness is not innate, and so neither can be the desire that Fa. This argument generalizes against any innate desire. Since content is determined only by causation and in triangulation, Hume was right about innate concepts, so this route is of no avail.

---

[143] ...at Darwall 1983, pp. 39-41.

She could posit the same moral desires, but all of them acquired. These desires would be basic because unrationalized, rather than basic because innate. The response to the Roberta case would be this. It's true that Roberta did not before have any relevant desire. Her new desire is the desire to relieve suffering, and it is not derived from any more basic desire. But that only counts against Hume's theory if the desire is nevertheless derived. But it isn't: it's not rationalized at all. It just so happens that, when we acquire the concept of suffering, we also acquire the desire to alleviate suffering. I have an argument against this suggestion, but it exactly matches my argument against the next suggestion, so I'll delay it a moment.

The Humean might posit an acquired generalized desire for the good. Such a claim could be either a contingent psychological generalization, or else a philosophical claim analyzing some other concept, like rationality. If the claim is a psychological generalization, then I don't think that it's very interesting and I can't imagine what the Humean could even begin to use as evidence for it. On the other hand, if the claim is intended to be something like a partial analysis of rationality, then I think that it has much merit but has lost touch with the spirit of Hume. The claim would be that it's partly constitutive of rationality that the rational agent desires the good. Perhaps there is an interpretation of Hume according to which that sort of claim is natural to Hume. But it gives the appearance of being a synthetic *a priori* claim of exactly the sort no Humean could accept.

Likewise if the posit is of many desires, such as a basic desire to ease suffering. Does the Humean intend a mere psychological generalization about people's responses to acquiring the concept of suffering? Or does she aim to tell us something about the constitution of the concept of suffering? If the former, the position is uninteresting. If the latter, it's not Humean.

To prepare for a future argument, let me return to the first Humean claim. Perhaps the Humean can compromise with the Kantian. The Humean can maintain that a desire in

a more-than-formal sense is always necessary for motivation, while the Kantian can insist that belief alone can motivate. This compromise can be carried off by the introduction of "besires," states that are both beliefs and desires.[144] I will argue that there could be no such states. Little offers a good treatment of what besires would be:

> ...there is no one propositional content of a mental state with doubled directions of fit, anymore than there is one propositional attitude involved in such a state. It is a state with two *complex* properties: it is a believing-attitude directed toward one proposition, and it is a desiring-attitude directed toward another. (Little 1997, p. 64)

Consider a randomly chosen belief and desire. I desire that there be ice in my water, and I believe that the Eiffel Tower is in Paris. This belief and this desire could not, even by coincidence, be the same attitude. We may assume that what attitude it is, and what its contents are, are essential to any attitude. So, if there were a bidirectional attitude, then it would be essentially the belief that it is, and also essentially the desire that it is. Thus, if it could do without either its belief component or its desire component, then that component is not, actually, a component of the attitude, since it can't be that attitude without having that component if that sort of component is essential to its identity. Assume that my desire for ice is my belief about the Eiffel Tower. But surely I could still desire ice even if I had no belief, or a different belief, about the location of the Eiffel Tower. Thus the desire for ice is not, actually, a belief about the Eiffel Tower.

Of course, that sort of example is not what compromisers have in mind. What they propose is that the belief that such-and-such would be a good thing for the agent to do is at the same time the desire that the agent do such-and-such. Even though this is much more plausible, it seems to me to fall prey to the same sort of problem. It overstates the strength of Kantian pure practical reason to say that, given some belief about what I ought to do, I *necessarily* — for identity is a relation that holds of necessity or not at all — desire to do it. Surely any (not 'every') moral belief that I have, I could have without

[144] This idea or at least the coinage is due to Altham, 1986; I rely on others' treatments.

the appropriate matching motivation. Likewise, for any (not 'every') moral motivation of mine, I could have it without thinking that the act in question is the right thing to do. The desire could have some other rationalization, or even be basic.[145]

The Humean can't compromise with the Kantian, at least not like that. The first Humean claim seems to be trivial, and the second controversial at best. Let's take another look at the first claim. Michael Smith has proposed a simple and well-known argument for it, and, notably from my point of view, he quotes Davidson when stating the thesis: "*R* is a primary reason [i.e., explanatory reason] why an agent performed the action *A* under the description *d* only if *R* consists of a pro attitude [i.e., desire] of the agent toward actions with a certain property, and a belief of the agent that *A*, under the description *d*, has that property." (Davidson 1963, p. 5) Davidson himself offers no argument for this claim at all; his concern is with the kinds of relations explanatory reasons have to action and reconciling the rationalization relation with the causal relation. Smith, though, does offer an argument with that thesis as its conclusion:

> (a) Having a motivating reason [i.e., explanatory reason] is, *inter alia*, having a goal.
> (b) Having a goal is being in a state with which the world must fit.
> (c) Being in a state with which the world must fit is a desire. (Smith 1994b, p. 116)

Like Ross, Smith appeals to desires' direction of fit to show that they are necessary for action. This argument is better than Ross's, which held that an attitude with the world-to-mind direction of fit had to descend from another such attitude. There's just no reason to believe that.

Smith has it that worldly action performed to satisfy some attitude, must be performed to satisfy an attitude with the sort of direction of fit that makes the world responsible for satisfying the attitude. This more abstract (than Smith's) version of the argument is sound. Noteworthy about the actual argument is that premise (c) is obviously

---

[145] This argument is similar to Smith's attack on besires at Smith 1994b, pp. 117-20. It is, of course, reminiscent of Descartes' and Kripke's arguments for mind-body dualism, but it doesn't make the mistake of their argument(s): treating apparent conceivability as indicative of metaphysical possibility.

false. Desires are not the only attitudes that have the world-to-mind direction of fit. Intentions do, too.

This isn't semantic fiddling; intentions are not a kind of desire, they are a distinct sort of attitude. Like desires, they have the world-to-mind direction of fit: when they aren't satisfied, it is the world, rather than the intentions, that are at fault. But intentions have different relevance. As Bratman notes,[146] intentions play a distinct role in practical reasoning. Whereas a desire can still exist despite being set aside or outweighed, when an intention is set aside, it no longer exists. As long as an agent genuinely intends to do something, that intention plays a role in practical reasoning: decisions must allow the intention to remain satisfiable. If the agent makes an intention of hers unsatisfiable, and knows that, then the intention no longer exists. A desire, however, can exist despite one's practical reasoning determining that it will not be satisfied. This settledness of intentions is important.

Davidson has an interesting suggestion about intentions and desires.[147] Unfortunately, his phrasing makes things hard, for he calls both desires and intentions 'judgments,' which they obviously aren't. Neither desiring nor intending to do something is believing that it is desirable. Some charity is called for: here is his idea. We often have contradictory desires. I begin with the desire, say, to lose weight, and also the desire to eat doughnuts. This is a doughnut and I can eat it; but not eating it will help me lose weight. So I want both to eat the doughnut and to refrain. It's not reasonable to perform an act just because it belongs to the class constituting the content of some desire of mine (even if there is only one member of the class). A desire is the desire to perform members of a class of actions, all other things being equal. Since all other things are rarely equal, the desire must be played off against other attitudes in practical reasoning. Assume that I decide to eat the doughnut. It's still true that it's unreasonable to eat this doughnut just

[146] See Bratman 1984.
[147] See Davidson 1978, pp. 96-9.

because I would like to, all other things being equal. It's reasonable to eat the doughnut because eating it seems to be the best thing to do, all things considered. To complement desires, which are all-other-things-being-equal attitudes with world-to-mind direction of fit, we need intentions, which are all-things-considered attitudes with world-to-mind direction of fit.

The point of this has been that Smith's premise (c) is wrong. We should agree that, whenever we act, our action is caused and rationalized by an attitude with the world-to-mind direction of fit. We need not agree that the attitude is a desire. Intentions will do just as well, and probably better in many cases. Intentions are, in Bond's term, mere "ghosts" of the moral judgments that somehow bring them about. That may account for some differences in phenomenology that there may be between things that we want to do and things that we do from a moral sense. It's also presupposed by the common claim that an agent performs an act, not because she wants to, but because she has to or thinks that she ought. Scanlon agrees about the motivational efficacy of intentions:

> ...a rational person who judges there to be compelling reason to A normally forms the intention to do A, and this judgment is sufficient explanation of that intention and of the agent's acting on it.... There is no need to invoke an additional form of motivation beyond the judgment and the reasons it recognizes, some further force to, as it were, get the limbs in motion. (Scanlon 1998, pp. 33-4)

Scanlon presages my approach, though I will reintroduce desires. What's pointed out in his remark is that moral judgment somehow translates into moral action by way of some attitude with the appropriate direction of fit. What we need to know is *how* judgment leads to the appropriate attitude.

Before turning to a few views on that subject, I want to mention the oft-quoted but terribly obvious contribution of Michael Stocker, that we are sometimes indifferent to what we agree to be good, and sometimes positively desire what we agree to be bad. Stocker notes that "only against a certain assumed background of agent mood and interest does citing the (believed) good make an act intelligible," (Stocker 1979, p. 746), and that "motivation and evaluation [i.e., moral belief, and desire or intention] need not point in

the same direction, that they are related through complex structures of mood, care, energy, interest, and the like." (*ibid.* p. 750) However we relate moral belief to moral desire or intention, we should bear in mind that the connections are defeasible and often complex. There might be some background condition required for beliefs to bring about intentions.

The background condition offered by several philosophers is rationality. My general line is offered by Christine Korsgaard, who says:

> In order for a theoretical argument or practical deliberation to have the status of reason, it must of course be capable of motivating or convincing a rational person, but it does not follow that it must at all times be capable of motivating or convincing any given individual. It may follow from the supposition that we are rational persons and the supposition that a given argument or deliberation is rational that, if we are not convinced or motivated, there must be some explanation of that failure. (Korsgaard 1986, p. 321)

For Korsgaard, a rational agent will be moved by moral judgment to perform moral actions. We've seen that action requires intention, so we may say that on Korsgaard's view, moral beliefs defeasibly bring about moral intentions. My concern in this section is with *how* that happens. What is rationality, such that it can effect this transition in the rational? And my answer will appeal to the holistic nature of the attitudes.

Velleman offers one sort of account, one that appeals to a central moral desire to, as a Humean would have it, serve as a derivation base for moral intentions and other moral desires. Velleman proposes that the premises of the argument for Humeanism goes like this:

> Suppose that reasons for someone to do something must be considerations that would sway him toward doing it if he entertained them rationally. And suppose that the only considerations capable of swaying someone toward an action are those which represent it as a way of attaining something that he wants, or would want once apprised of its attainability. (Velleman 1996, p. 170)

This argument appears to be one about the derivation of desire. A new desire can be generated from new beliefs only if those beliefs are about ways of satisfying old desires. What Velleman wants us to notice about the argument is that:

> The first premise... doesn't entail that if a consideration fails to influence someone, then it isn't a reason for him to act; it entails that if a consideration fails to influence someone, then either it isn't a reason for him to act or he hasn't entertained it rationally. The inclinations that would make an agent susceptible to the influence of some consideration may therefore be necessary — not to the consideration's being a reason for him — but rather to his being rational in entertaining that reason. (*ibid*, p. 172)

For Velleman, the rational consideration of certain facts is itself adequate to account for moral motivation. That's because certain inclinations are intrinsic to rationality. The inclination to which Velleman appeals[148] is an inclination toward autonomy. I think that Velleman's account of this inclination is unintelligible or absurd. But I want to generalize. Velleman's view consists, for my purposes, of two claims. First, that the Humean is right to say that desire is only derived from desire. Second, that rationality is partly constituted by the possession of some core moral desire. The second claim neuters the first: that's why I earlier suggested that this sort of view was not Humean in spirit.

Is the second claim plausible? Michael Smith has an argument against claims of this sort:

> ...the objection... is... that, in taking it that a good person is motivated to do what she believes right, where this is read *de dicto* and not *de re*, externalists... provide the morally good person with 'one thought too many'. They alienate her from the ends at which morality properly aims. ...it is constitutive of being a morally good person that you have direct concern for what you think is right, where this is read *de re* and not *de dicto*.
>
> [An explanation like Velleman's] elevates a moral fetish into the one and only moral virtue. (Smith 1994b, p. 76)

Smith is attacking externalists about motivation who claim that a desire to do the right is what accounts for derived moral desires. Since Velleman's view about motivation similarly posits one core desire from which all other moral desires are derived, it's formally like the externalist views whose form Smith is attacking, so Smith's argument ought to work against Velleman if it works at all.

---

[148] Velleman 1996, pp. 193-8.

Smith's argument concludes that we don't have a desire for the good; rather, we have desires for all of the particular goods. Attribution of desires and intentions, like attributions of belief, occurs in radical interpretation. We approach an agent with a synthetic *a priori* theory of interpretation, and we attempt to understand her actions in the terms of that theory. When an agent acts, we must attribute some attitude to rationalize and cause that act: an intention or perhaps a desire. To respect the holism of the attitudes, we must then also attribute other relevant beliefs, desires, and intentions in an intelligible network. What hermeneutical considerations would lead us to posit a desire to do the right or achieve the good? No such desire, Smith might say, is necessary to account for action. And it isn't obvious why one would be.

Smith has his own model whereby rationality makes for the transition from belief to desire. Smith speaks of two relations between moral beliefs and moral desires. First, there is a normative relation captured in his thesis C2: "If an agent believes that she has a normative reason to ϕ, then she rationally should desire to ϕ." (Smith 1994b, p. 148) But the relation is not only normative: beliefs about normative reasons can cause desires.[149] Because the belief makes it rational to have the desire, the belief's actually bringing about the desire counts as a rationalization. So far, the account is plausible.

The trouble comes in with Smith's account of normative reasons. Characterizing Korsgaard's view as a "platitude," Smith tries to clarify what the platitude says:

> ...it tells us that what it is desirable for us to do in certain circumstances — let's call these circumstances the 'evaluated possible world' — is what we, not as we actually are, but as we would be in a possible world in which we are fully rational — let's call this the 'evaluating possible world' — would want ourselves to do in those circumstances.
>
> In terms of the background idea, facts about what it is desirable for us to do are constituted by the facts about what we would advise ourselves to do if we were perfectly placed to give ourselves advice. (*ibid*, pp. 151-2)

[149] Smith 1994b, p. 179.

I have a rational self, and it has desires for me. Those desires are, *ex hypothesi*, rational. Since my rational self is the one with the desires, then surely I should share them, and, if I don't, then I'm irrational. My reasons are constituted by the desires of my rational self.

The problem should be obvious. Given a rational self with a set of *ex hypothesi* rational desires, it's obvious that, if I don't have those desires, then I'm irrational. But how did my rational self get the desires? What was it about his rationality that led him from moral belief to moral desire? Smith's account was to explain how rationality moves me from belief to desire, but it relies on my prior agreement that my rational self has rational desires. But to understand that, I have to understand what is rational about my rational self's desires. Smith's account is tantalizingly close to right, but it seems to give up at the crucial moment.

I want to lay out a model of moral motivation. I need to begin by drawing a distinction between two kinds of defeasible connectives to which I appeal: rationally defeasible and causally defeasible connectives. A rationally defeasible connective appears in a defeasible principle that we attribute to someone as the content of her belief. When we attribute defeasible principles, we attribute an attitude that, while being completely rational, only defeasibly leads the thinker from premises matching the left of the principle to a conclusion matching the right.

A causally defeasible connective appears in an hermeneutical rule that we use to interpret agents. When we attribute someone an attitude, we implicate many other attitudes that, all other things being equal, she has. Were she fully rational, she would actually have all of these other attitudes. But full rationality is an ideal never empirically realized. On the other hand, full rationality is that with reference to which we define rationality, so, to be rational, even less-than-fully, one must approximate to full rationality. Thus one must, all other things being equal, have the attitudes that one would have if fully rational. We should therefore approach interpretation with such causally defeasible rules in mind. While attributing the attitudes on the left only defeasibly implies

that we should attribute the attitudes on the right, that defeasible connection is what constitutes the attitude, and the agent, as rational.

With this distinction in mind, I can offer two sets of causally defeasible hermeneutical rules about moral judgments and motivation. Let me begin with thin directive judgments, judgments like that this would be the right or best thing to do, or is the thing that I ought to do. I want to suggest that it's partly constitutive of judgments like this that they are defeasibly connected with moral intentions. My argument, predictably, is hermeneutical. For imagine that an agent routinely said that such-and-such would be the right thing for her to do right now, all things considered, but rarely or never did such-and-such. Whatever role her concept 'right' plays in moral reasoning is entirely unlike the role played by our concept 'right' in ours. Thus, by the anti-relativist argument, it is a different concept.

The argument is that thin directive judgments must be connected with action, but the conclusion is that they're connected with intentions. That is not a logical gap, for action is connected with intentions. Intentions to act are logical ghosts of action. They stand, hermeneutically, between the belief that this is the right action, and this action. Their role is to cause and rationalize action in a way that belief can't, since belief has the wrong direction of fit.

The alternative concept of 'right' that I mention in the argument might be co-extensional with our concept. But, as I've tried to argue, the identification of the content of attitudes with external truth-conditions shows that those contents are opaque. Two attitudes might share content, but have different logical and motivational connections with other attitudes, because, though they are the same sort of attitude toward the same content, they understand that content differently. Moral judgments are attributed on the basis, not just of their truth-conditions, but also of their place in the whole mental life and agency of the moral judge.

On the other hand, to count as an agent, one must perform intentional acts, and an intentional act is typically one caused and rationalized by an intention that had that act as its content. But, I suggest, an intention can only be understood when attached to a belief that its content is, all things considered, the best or right thing to do. Agents as such must have these thin moral concepts. It takes interpretation to find out how they express them verbally, and it might not be in the same terms that we use, but these facts are a commonplace.

While thin directive judgments are to be associated with intentions, thick directive judgments are to be associated with desires. Just as being generous defeasibly (all-other-things-being-equal) implies being the right thing to do, likewise one would like, defeasibly (all-other-things-being-equal), to do the generous thing. Here, appeal to my standard hermeneutical consideration is troubled. I suggest that thick judgment defeasibly implies desire. But that desire is one among many, none of which issues directly in action. Being motivated to perform an act is partly constitutive of believing that it would be generous, but performing an act is not partly constitutive of being motivated to perform it. So I can't just take the normal case of action and offer an interpretation in the form of attribution of moral desires. The argument that thick moral judgments are connected to desires will have to take an alternate route.

I want to describe two situations, the normal situation, and one kind of abnormal one. In the normal situation, desires play no role. An agent has a number of beliefs, thick moral judgments and other, non-moral but normatively relevant beliefs, which she considers. She engages in practical reasoning, considering which features of possible actions are most right-making, which features undermine other features' right-makingness, which features have which place in various hierarchies, and trying to do what she thinks, all things considered, to be best. She comes to a conclusion, a thin directive judgment that *this* is what she ought to do, or would be right or most good.

Since it's constitutive of such judgments that they defeasibly cause intentions to perform accordingly, the practical reasoning issues in belief, intention, and hence action.

Note that the following sort of situation would be epistemically odd, but not abnormal from the point of view of moral motivation: an agent judges that an action would be right, even though she does not apply any positive thick moral concept to the act. The act is neither generous nor just nor anything else of the kind. This sort of situation can emerge because the thin moral concepts outstrip any statable principles we could use to derive them from thick moral concepts. Where this situation emerges, no moral desires will play any role at all in practical reasoning or moral motivation.

The abnormal case gives us our reason for attributing desires to be generous, just, and the rest. The abnormal case is a kind of moralistic *akrasia*. Here is what Davidson has to say about *akrasia* in general:

> Why would anyone ever perform an action when he thought that, everything considered, another action would be better? If this is a request for a psychological explanation, then the answers will no doubt refer to the interesting phenomena familiar from most discussions of incontinence: self-deception, overpowering desires, lack of imagination, and the rest. But if the question is read, what is the agent's reason for doing *a* when he believes it would be better, all things considered, to do another thing, then the answer must be: for this, the agent has no reason. ...in the case of incontinence, the attempt to read reason into behaviour is necessarily subject to a degree of frustration. (Davidson 1970b, p. 42)

Davidson's two ways of reading the question may be correlated to two descriptions of the action in question. Assume that the agent wanted to eat a doughnut, but decided that, all things considered, she shouldn't. And yet she finds herself driving to Krispy Kreme. If we described the action as eating a doughnut, then we can give an account for it. Its rationalizing cause is the desire to eat a doughnut.[150] If we describe the action as doing what she thought, all things considered, was not the best thing to do, then there is no rationalization to be given. Akratic action, I would say, is action whose rationalizing cause is desire, rather than intention, and which is such that one's intentions are being

[150] This sort of case is why I said that intentional action "typically" flows from intentions.

thwarted by the action. This line commits me to the oddness of having to say that some intentional actions (driving to Krispy Kreme is not, in the example, something done by accident) are not caused by intentions, but I accept the consequence. The connection between intentions and intentional actions is only defeasible, but adequate to account for the conceptual and verbal link between the two.

The point of the account is this. Akratic action is action, and it deserves a rationalizing cause. However, it does not deserve an intention as cause, because an intention is an all-things-considered attitude, and akratic action is necessarily not what the agent things it's best, all things considered, to do. So we must find an attitude that has the appropriate direction of fit to cause, in a rationalizing way, akratic action: that's desire. Furthermore, being caused by desire is not adequate: akratic action must frustrate one's all-things-considered judgments, which themselves yield intentions. So akratic action is action caused by desires in contravention of one's intentions (or at least the intentions that one's thin directive judgments would come attached to, if the whole process weren't being defeated by the akratic action).

The account of *akrasia* appeals to desires. Typically, one performs an akratic action for non-moral reasons. But surely, a person can believe that an act, while generous, is not the best thing to do all things considered, and yet akratically perform it anyway. If such moralistic *akrasia* is possible, then there must be moral desires that flow from thick directive judgments. Such desires will play the same role in moral reasoning that desires play in other sorts of practical reasoning: usually, no role at all, but sometimes, they override intention and lead to *akrasia*.

The core of the account is the connection between thin directive judgments and intentions; moral desires are posited only because of the possibility of moralistic *akrasia*, and if that sort of action is impossible, then we can do without moral desires entirely. This account resembles that of R. Jay Wallace[151]; Wallace speaks of "association"

[151] See Wallace, 1990, esp. pp. 363-6.

between desires and evaluative beliefs, where I speak of causation, but I think that he has causation in mind. For Wallace, desires are rationalized by way of their associated evaluative beliefs being rationalized, but I suggest that the causal relation between thick directive judgment and desire, and thin directive judgment and intention, is itself a relation requiring rationalization. Luckily, possessing such rationalizing, causally defeasible, causal relations is partly constitutive of moral judgments.

I said that Korsgaard's account pointed the way for mine. In what way have I appealed to rationality to move the agent from belief to action? My account appeals to the rationality of the agent, in the form of her being interpretable despite holistic restrictions on interpretation. To be interpreted as having the moral belief that thus-and-such would be the right thing to do, an agent must actually perform thus-and-such most of the time. To do otherwise is not irrational but arational.

I said that Smith's account was nearly right. He appeals to rationality to account for the shift from belief to desire. In what way is my account different from his? I merely add that having the causal power to inspire intention is *partly constitutive* of moral beliefs. Moral beliefs are attitudes that we cannot attribute to an agent without also attributing to her the intention to perform an appropriate action. She (defeasibly) cannot have moral beliefs without having intentions to perform accordingly. That is the role of rationality: it provides a holistic constraint on the possession of moral attitudes.

5.2.5 Comparison of New Wave and Hermeneutical Moral Realism

Hermeneutical and New Wave Moral Realisms have similar, but not identical, metaphysical foundations. They are, however, very different in detail. In this section, I want to discuss the similarities and differences between Hermeneutical Moral Realism (hereafter, HMR) and its closest cousin (hereafter, NW).

First, I want to discuss their relationships to the Open Question argument. I discussed NW's reply in 5.1. Moral properties may be identified synthetically, since

meanings are worldly; hence Moore's demand that moral properties be identified analytically or not at all is a mistake. How can HMR respond?

Not the same way, and this leads into the main difference between HMR and NW. NW is based on a term-by-term theory of language. A certain object causes the use of a certain name, and a certain property causes the use of a certain predicate; thus that object and property are the referents of the name and predicate respectively. On my Davidsonian approach, words aren't the units of meaning, sentences are. Truth-conditions are the causes relevant for grounding meaning in the world. HMR doesn't accept moral properties any more than its semantic and metaphysical bases accept any kind of properties. Thus HMR rejects the demand that goodness be identified at all.

No doubt a suitably revised version of the Open Question Argument could be developed having to do with predicates rather than properties. In that case, HMR would again reject the demand that satisfying moral predicates like '...is good' can be identified as satisfying some other set of predicates. Rather, the thin moral concepts are anomalous with respect to non-moral concepts; they stretch beyond any statable principles relating the two. For Moore, goodness must be absolutely simple, since it's undefinable; for HMR, 'good's indefinability is indicative of its complexity.

Thus HMR lacks the first two components of NW that I discussed in 5.1, synthetic statements of property identity and rigid designation. There are no properties to be identified and no designation, rigid or otherwise. Kripke's and Putnam's semantic insights took place in the context of a theory of meaning that asked after the meaning of a word out of the context of a sentence, and so can only inspire, rather than instruct, semantic theory.

NW and HMR also have similar analogies with philosophy of mind. NW looked to (Aristotelian?) functionalism's approach to mental properties for an analogy for moral properties. HMR looks to Davidson's Kantian anomalous monism for the analogically anomalous properties. Functionalism and anomalous monism are alike in being token-

identity theories. Functionalism, however, looks to a fundamentally biological approach to the mind, while anomalous monism is resolutely dualistic in its approach the mental concepts. To plumb the depths of this difference would lead me too far astray; let me say only that a version of moral realism that finds a home in a biologistic conception of rationality threatens to overwhelm human agency and autonomy — our spirituality, one might say — with physical criteria masquerading as moral.

If we focus on the idea of multiple realizability, though, we find more harmony. In functionalist approaches to mind, multiple realizability is the idea that functionally, and hence mentally, identical states can be realized in a multitude of physical substrata. So in NW, the idea is that moral properties, which are functionally defined, are multiply realizable. HMR would agree, so long as the point is rephrased without reference to properties.

Moral concepts are related to one another, and to non-moral concepts, by way of principles. However, the left sides of the principles are unstatably complex disjunctions. Each disjunct states a possible satisfaction condition for the concept on the right. Thus there are multiple possible realizations of the concept on the right, one for each disjunct.

Likewise, the naturalistic and coherentist epistemology of NW is not alien to HMR: HMR's moral epistemology is the pursuit of reflective equilibrium. However, we will not, in reflective equilibrium, identify the properties that have been causally connected with our use of moral language, as there are no such properties. Rather, each moral belief, once embedded in a true Tarskian truth theory for the speaker, already identifies its truth-condition, and, if a moral truth-condition can stand as the cause of a belief, then such beliefs identify their causes. The difference here is based on the fact that NW seeks to synthetically identify moral properties. For HMR, moral concepts' relations to other concepts resist codification because of their great complexity and their anomalism with respect to the non-moral.

The final difference is with respect to moral motivation. NW has externalistic aspirations: it wants the relation between moral belief and moral motivation to be an external one. HMR makes the relation defeasible, but internal. It's part and parcel of being a moral belief that an attitude cause appropriate desires and intentions.

5.3 MORAL SCEPTICISM

The *locus classicus* for moral scepticism is, of course, J. L. Mackie's *Ethics: Inventing Right and Wrong*. I want to address the main points of Mackie's scepticism before moving on to the Moral Twin Earth argument, which is intended to be a version of one of Mackie's arguments.

Mackie calls his first argument the Argument from Relativity. This argument is an inference to the best explanation. There is a diversity of moral beliefs. What can best account for this? Mackie says:

> Disagreement about moral codes seems to reflect people's adherences to and participation in different ways of life. ...the argument from relativity has some force simply because the actual variations in the moral codes are more readily explained by the hypothesis that they reflect ways of life than by the hypothesis that they express perceptions, most of them seriously inadequate and badly distorted, of objective values. (Mackie 1977, pp. 36-7)

If adherence to a way of life amounts to believing a certain moral theory, then part of Mackie's remarks amounts to explaining difference with reference to difference. He's on stronger ground with 'participation.' Again, his use of 'perception' is troubled. If Mackie seriously means to say that, for the moral realist, moral properties are phenomenally given to us just like perceptual properties, and that the moral realist must appeal to some sort of moral hallucination to account for moral error, then of course his moral realist opponent has nothing to do with hermeneutical moral realism. We should broaden the idea to moral belief. Mackie's more plausible claim, then, is that differences in moral belief are best accounted for with reference to non-cognitive determination of moral beliefs by ways of life, rather than with reference to cognitive error.

It's not plain, though, why the moral realist must choose. Because of the veridical nature of belief, true belief can pass unremarked, but error requires an account, and where there is moral disagreement, there is moral error. The moral realist must look somewhere to account for moral error. Why not look to culture? As culturally embedded people, we learn morality, but we also learn a set of interests and motivations that distort our moral reasoning. The distortions can be made apparent in the pursuit of reflective equilibrium.

The argument from relativity can gain no purchase, because no form of conceptual relativity can gain purchase. To count as the same beliefs, two beliefs must have identical truth-conditions. Conceptual relativity requires that there be beliefs with the same meaning but different truth-values. Obviously, on the truth-conditional account of meaning, that's quite impossible.

Mackie's second argument is the Argument from Queerness, and it has two parts, a metaphysical part and an epistemological part. I want to focus on the epistemological part first. Mackie says that "...if we were aware of them [objective moral values], it would have to be by some special faculty of moral perception or intuition, utterly different from our ordinary ways of knowing everything else." (*ibid*, p. 38) I just don't know why this is supposed to be true, but I suspect that Mackie is taking for granted a foundationalist approach to knowledge in which the foundations are justified by perceptual experience. Such an approach is incoherent, for experience stands in no epistemic relationship to belief.

Justified moral beliefs can come about in one of two ways. They could be caused by their truth-conditions; I'm sceptical of such an account, but I would be delighted if it were true. Or they could be caused, in a rationalizing way, by reasoning from other, non-moral beliefs, according to defeasible synthetic *a priori* moral principles. Such principles are no more "queer" than are principles governing the application of intensional concepts, and they are not a faculty of perception or intuition. Just as the argument from relativity

presupposed internalism about content, this epistemic argument from queerness presupposes empirical foundationalism about epistemology.

The other argument from queerness has to do with moral motivation:

> If there were objective values, then they would be entities or qualities or relations of a very strange sort, utterly different from anything else in the universe.... An objective good would be sought by anyone who was acquainted with it, not because of any contingent fact that this person, or every person, is so constituted that he desires this end, but just because the end has to-be-pursuedness somehow built into it. (*ibid*, pp. 38, 40)

Here, Mackie doesn't rest on some assumption that I can quickly label and dispense with: it's the labeling, though, and not the dispensing, that's the problem. I claim that moral beliefs motivate. However, a belief could have the same truth-condition as a moral belief, without being a moral belief. Moral facts thus do not have to-be-pursuedness somehow built into them. Moral beliefs have to-be-motivating built into them, as a constitutive fact about such beliefs. Neither contingent desire nor external fact accounts for moral motivation: the holistic nature of moral belief does.

This quick tour of Mackie's scepticism can't delve much into the details, but I hope to have made the basic point that traditional moral scepticism finds life in the context of internalism and foundationalism. Moral scepticism requires retooling to deal with contemporary forms of moral realism. My main target in this section is just such a retooled moral scepticism, the Moral Twin Earth argument of Horgan and Timmons. The argument can be seen as a generalization of an argument offered by Hare.[152] Hare offers the example of a missionary among cannibals. The missionary and the cannibals obviously have a moral disagreement of some kind: the missionary opposes the cannibals' cannibalism, while the cannibals favor it. This disagreement obviously requires shared content. But if content is to be descriptions associated by speakers with their terms (by Fregean senses), then the missionary and the cannibals don't seem to share content. So they can't disagree; yet they do. So the content of normative concepts is not Fregean

[152] ...at Hare 1952, pp. 148-50.

senses. Hare concludes that a form of expressivism is correct, by a bad disjunctive syllogism in which the disjunction doesn't state all of the options. New Wave Moral Realism makes the meaning of normative concepts into worldly properties, not necessarily known to the speakers. So the moral disagreement between the missionary and the cannibals can be made possible by the properties to which they both refer with their normative terms, even though one or the other party is widely mistaken about the nature of those properties (and even which properties they are, perhaps). By negating the transparency of meaning, the semantic externalism on which the New Wave was based seems to solve Hare's problem.

Enter Moral Twin Earth, which, as Merli points out,[153] amounts to an updating or generalization of Hare's argument. Horgan and Timmons ask us[154] to imagine a Moral Twin Earth. Moral Twin Earth relates to Earth somewhat as Twin Earth does in Putnam's thought-experiment. For Putnam, what causally regulates our use of the term 'water' is water, whereas what causally regulates our Twins' use of the term 'water' is something else, twin-water. Thus our terms have different meanings. Since the Boyd-Brink view has it that moral terms are natural kind terms, the same sort of thought-experiment should distinguish our moral terms from those of Twins whose moral thought is causally governed by different natural properties. But it can't, so Boyd and Brink are wrong about the semantics of moral terms.

Here's the idea. We on Earth are gradually moving toward a reflective equilibrium in which, as Boyd and Brink say, we will identify the properties that have always governed our use of moral language. Assume that our reflective equilibrium will be some very sophisticated consequentialist theory. Note that the most sophisticated variants of deontology and utilitarianism offer the same judgments in most cases; we can assume that yet more sophisticated variants will move even closer together (though without ever

---

[153] ...at Merli 2002, p. 209.

[154] ...at Horgan and Timmons 1990, pp. 458-61, and Horgan and Timmons 1992, pp. 244-6.

agreeing on all cases). Our Twins on Twin Earth are also gradually moving toward a reflective equilibrium, but their equilibrium is a sophisticated deontological theory. Because the theories are in equilibrium, they will have massive overlap with one another, though without agreeing on every point.

Assume then that we meet our Twins. While our philosophers and theirs begin to wrangle with semantic and moral issues, our psychologists and anthropologists and theirs note that, while we are slightly more sympathetic in nature than our Twins, they are slightly more guilt-prone than we are. All of these differences are intended to be narrow enough to allow Moral Twin Earth to count as a Twin, rather than as freakishly alien.

As with the missionary and the cannibals, it's obvious that, in the marginal cases where our equilibria don't overlap, we have moral disagreement with our Twins. So we must share meaning. But if meaning is causally regulating property, then we don't, since our moral terms are governed by consequentialist properties, and theirs are governed by deontological properties. So we're faced with two alternatives. We could adopt relativism, and allow that we will never come to agree with our Twins about what to do, but that no one is wrong for all of that; or, we could give up on thinking that stating truths is something that moral judgments are any good at or are even designed for. Our options, then, are a perverse form of relativistic realism, or else expressivism. Neither option is acceptable to the realist. However, since neither option is acceptable period, realism is not the problem. There must be something wrong with the thought experiment.

Much discussion of the thought experiment has missed the point. For instance, Copp contends that "despite the fact that corresponding moral and twin-moral terms do not express the same property... moral terms might be the best *translation* for the corresponding twin-moral terms." (Copp 2000, p. 121) Since that's a premise of the argument, it's not obvious how it's supposed to help. Copp's contention is that we can disagree with our Twins because we share prescriptive, rather than referential, content

with them. But that is the conclusion of the Twin Earth argument, not at all an implication of New Wave Moral Realism.

David Brink offers a more interesting but no more helpful response. Brink gives up on what I see as the linchpin of the Boyd-Brink view, the notion of functional properties causally regulating our use of moral terms. He says:

> On this view, a natural property N causally regulates a speaker's use of moral term 'M' just in case his use of 'M; would be dependent on his belief that something is N, were his beliefs in dialectical equilibrium. ...worries [about New Wave Moral Realism] are less compelling when we shift from the extensional to the counterfactual or dialectical understanding of causal regulation. (Brink 2001, p. 169)

Brink's strategy is to reconceive the relation between moral property and moral term. If having the moral property causally regulate the use of the appropriate moral term as water does 'water' leads to problems on Twin Earth, then the causal regulation must take some other form.

"Extensional" causal regulation is the regulation that water does of 'water.' It's not at all obvious that "counterfactual" causal relation is any different. What Brink means by counterfactual causal regulation is causal regulation that is not defeated by any local, contingent error; the counter-to-fact hypothesis is that the speaker is in reflective equilibrium, and hence not at all error-prone. But surely the same applies to non-moral properties. The way in which water causally regulates 'water' surely accommodates the fact that we occasionally misidentify water. It's only our error-free identifications that we allow to count in determining the referent of 'water.' Likewise only our error-free identifications of moral properties count in determining the referents of the appropriate moral terms.

"Dialectical" causal regulation doesn't appear to be a species of causal regulation at all. Boyd's idea was that, in reflective equilibrium, we identify the properties that have always causally regulated our use of moral terms. "Dialectical" causal regulation, then, would be parasitical on "extensional" causal regulation. If we cut the dialectic loose from

the causal histories that determine the referents of our moral language, then the properties identified in reflective equilibrium need have had no particular causal relationship with our use of moral terms. The entire attraction of the New Wave was its appropriation of externalist semantics to answer the Open Question argument; the Kripke-Putnam semantics shorn of rigid designation of natural properties determined by causal histories would be a thin theory indeed.

Merli offers another incorrect defense. Merli notes that Horgan and Timmons accept that all Earth moral discourse is governed by the same set of moral properties; only we and our Twins show a difference. He then argues that

> The realist can preserve meaning across actual-world [i.e., Earth] disputes, we're assuming, though, if Horgan and Timmons are right, she *can't* preserve meaning between Earth and Moral Twin Earth. Thus the twin-moral practice must be different from our own moralizing in a way that puts it beyond the realist's reach. So there's pressure on Horgan and Timmons to make the details of twin-moralizing *much* different from those of our own moral discourse; in an important sense, it must be like nothing we've ever seen before. (Merli 2002, p. 218)

I think that this is a mistake. We use water for a wide variety of purposes. Our water-discourse has massive overlap with our twins' twin-water-discourse in the following sense: if we remove 'water' from many sentences, and replace it with 'twin-water,' we preserve truth. Only in some chemical and physical theories, and some very obscure uses, is there a difference. Likewise with goodness and whatever our twins refer to with 'good,' twin-goodness. If we remove 'good' from many sentences and replace it with 'twin-good,' we preserve truth. Only in some ethical theories, and some very obscure cases, is there a difference. But just as water is not twin-water, goodness is not twin-goodness.

I believe that the Moral Twin Earth argument is lethal to New Wave Moral Realism. Why is this relevant? New Wave Moral Realism is the closest philosophical relative to the Hermeneutical Moral Realism I offer. They both take off from externalist semantic theories, find analogues for morality in token-identity theories of the mental, and adopt coherentist methods in moral epistemology. Moral Twin Earth, then, is a

ready-made objection to Hermeneutical Moral Realism, one that I can use to show the strengths of the view. Naturally enough, the differences between the externalisms and token-identity theories that New Wave and Hermeneutical moral realisms are based on will account for their different prospects in the face of Moral Twin Earth.

The key difference is that the New Wave's externalism has to do with reference. Names and natural kind terms are held to have referents determined by causal histories of uses of the terms. For Davidson, reference is a theoretical construct. Kripke-Putnam style externalism is based on a term-by-term semantics that ignores Frege's dictum that we should look for the meaning of a word only in the context of a sentence. Such a theory will have a terrible time accounting for the unity of the proposition. Note that it's the causal regulation of moral terms that Horgan and Timmons finds objectionable. Hermeneutical Moral Realism does not accept that the causal regulation of moral terms gives them content; terms never have content at all, they only have systematic contributions to the truth-conditions of the sentences in which they appear.

The point of the Moral Twin Earth argument is that our twins have the same moral beliefs that we do — like us, they believe that things are good or wrong — but that their beliefs have nevertheless different truth-conditions and indeed different truth-values. That would show either that moral beliefs are externally related to their truth-conditions — relativism — or that they never had any truth-conditions at all — expressivism.

How do we identify moral beliefs? Hermeneutical Moral Realism identifies moral beliefs with reference to three relata: their truth-conditions, other beliefs with which they have inferential relations, and other attitudes with which they have rationalizing causal relations. Moral Twin Earth is a counterexample to Hermeneutical Moral Realism just in case the Hermeneutical Moral Realist must agree that our twins share our moral beliefs — like us, they believe that things are good or wrong — but also must agree that our twins' moral beliefs don't share our moral beliefs' truth-conditions. We should agree that our twins share our moral beliefs. The strategy is to move from that premise through the

commitments of Hermeneutical Moral Realism to the conclusion that our twins' beliefs must share our beliefs' truth-conditions.

Why do our twins share our beliefs? First, their moral beliefs have the same practical role that ours do. Like ours, their moral beliefs are internally, but defeasibly, connected with moral desires and intentions. Furthermore, and this is key, most of their moral beliefs are true under just the same circumstances that ours are. There is massive, though not total, overlap between their moral beliefs and ours. That was a condition of the thought experiment. Finally, there is substantial, though less, overlap between moral principles. For the sophisticated consequentialist as well as the sophisticated deontologist, that something is courageous or generous counts in favor of doing it, as a rule.

The contention of the Moral Twin Earth argument is that our twins have a distinct reflective equilibrium point from the one at which we aim. That could well be true, as long as consider only narrow reflective equilibrium. Narrow reflective equilibrium is the point at which my judgments and principles are fully coherent. If the truth-conditions of our moral beliefs are what we think they are in narrow reflective equilibrium, and we and our twins can have different narrow reflective equilibria, then we and they can have the same moral beliefs but with different truth-conditions. In that case, the argument has succeeded. I reject the premise that the truth-conditions of our moral beliefs are what we believe them to be in *narrow* reflective equilibrium.

First, wide, not narrow, reflective equilibrium represents the point of highest coherence in our moral beliefs. Our encounter with our twins would spur our philosophers and theirs to argumentative efforts on behalf of our respective divergent moral beliefs. Recall, specifically, that Horgan and Timmons offered to account for the difference in moral theories with reference to a difference in moral psychology: we're more sympathetic, our twins are more guilt-ridden. Perhaps the form of moral debate that we would engage in with our twins would be one in which we tried to convince each other that our moral psychologies were slightly erroneous in emphases. Sympathy and

guilt are holistic attitudes just like all the rest, and they have inferential and other relations with other attitudes; they can be changed by changes elsewhere in the system. Only after extensive debate with our twins would they and we enter into a state of wide reflective equilibrium, and there's no reason at all to believe that we would enter into different ones. Indeed, it would simply beg the question against the realist to assert that rational moral enquirers can't end at the same point because of contingent moral psychology. The truths of morality, I claim, are synthetic *a priori*: not contingent, but constitutive truths governing the application of moral concepts to the world. Since we agree that our twins share our concepts, then their correct applications and ours must be the same.[155]

Second, even wide reflective equilibrium doesn't represent a stable stopping point. Our moral concepts outstrip in complexity our ability to articulate them in statable moral principles. Moral enquiry, like all philosophical enquiry, is a never-ending, terminally open, process. Moral realism is as always a form of realism; we shouldn't look at any epistemic process to determine, but only to discover, the truth. Even if we never achieved complete agreement with our twins, asymptotic approach to agreement across time would indicate common truth-conditions. Common truth-conditions, not agreement, is what we need to defend moral realism. Moral Twin Earth can't force us to divorce moral meaning from moral truth-conditions, so it fails against Hermeneutical Moral Realism. I conclude that the truth-conditional theory of meaning shows that some moral beliefs and utterances are true or false, and that some of those are true.

[155] This reply is basically the same as Merli's, only without mistaken reference to properties.

References

Allison, Henry E. 1983. *Kant's Transcendental Idealism An Interpretation and Defense*. New Haven and London: Yale UP.

Altham, J.E.J. 1986. "The Legacy of Emotivism." in McDonald, Graham, and Wright, Crispin. 1986. *Fact, Science, and Morality: Essays on A.J. Ayer's Language, Truth, and Logic*. Oxford: Basil Blackwell. pp. 275-88.

Ayer, Alfred Jules. 1946. *Language, Truth, and Logic. 2nd Edition.* New York: Dover.

Audi, Robert. 1986. "Acting for Reasons." in Mele, ed. 1997. pp. 75-105.

Bar-On, Dorit. 1994. "Conceptual Relativism and Translation." in Preyer, Gerhard; Siebelt, Frank; and Ulfig, Alexander, eds. *Language, Mind, and Epistemology*. 1994. Dordrecht: Kluwer. pp. 145-70.

Beaney, Michael, ed. 1997. *The Frege Reader*. Oxford: Blackwell Publishers.

Bennett, Jonathan. 1988. *Events and their Names*. Indianapolis: Hackett Publishing Co.

Berkeley, George. 1734. *A Treatise Concerning the Principles of Human Knowledge*. in Ayers, M.R., ed. 1989. *Berkeley Philosophical Works including the works on vision*. London: Everyman's Library.

Bernstein, Richard. 1983. *Beyond Objectivism and Relativism Science, Hermeneutics, and Praxis*. Philadelphia: University of Pennsylvania Press.

Bittner, Rüdiger. 2001. *Doing Things for Reasons*. Oxford: Oxford UP.

Björnsson, G. 2001. "Why Emotivists Love Inconsistency." Philosophical Studies 104 (2001). pp. 81-108.

Blackburn, Simon. 1981. "Reply: Rule-Following and Moral Realism." in Holtzmann, Steven, and Leich, Christopher, eds. 1981. *Wittgenstein: To Follow a Rule*. London: Routledge. pp. 163-87.

ibid. 1984. *Spreading the Word Groundings in the Philosophy of Language*. Oxford: Clarendon Press.

ibid. 1985. "Errors and the Phenomenology of Value." in Blackburn, 1993a. pp. 149-65.

ibid. 1988a. "How To Be An Ethical Anti-Realist." in Blackburn, 1993a. pp. 166-81.

ibid. 1988b. "Attitudes and Contents." in Blackburn, 1993a. pp. 182-97.

ibid. 1990. "Just Causes." in Blackburn 1993a, pp. 198-209.

ibid. 1992. "Gibbard on Normative Logic." Philosophy and Phenomenological Research LII:4 (December 1992). pp. 947-52.

ibid. 1993a. *Essays in Quasi-Realism*. Oxford: Oxford UP.

ibid. 1993b. "Realism: Quasi, or Queasy?" in Haldane, John, and Wright, Crispin, eds. 1993. pp. 365-84.

ibid. 1998. *Ruling Passions A Theory of Practical Reasoning*. Oxford: Oxford UP.

Boghossian, Paul. 1989. "Content and Self-Knowledge." in Ludlow, Peter, and Martin, Norah. 1998. *Externalism and Self-Knowledge*. Stanford: CSLI Publications. pp. 149-74.

ibid. 1990. "The Status of Content." Philosophical Review XCIX:2 (April 1990). pp. 157-184.

ibid. 1998. "What the Externalist Can Know A Priori." in Wright, Smith, and Macdonald, eds. 1998. pp. 271-84.

Bond, E.G. 1983. *Reason and Value*. Cambridge: Cambridge UP.

Bonevac, Daniel. 2003. *Deduction Introductory Symbolic Logic 2nd ed.* Malden: Blackwell Publishing.

BonJour, Laurence. 1985. *The Structure of Empirical Knowledge*. Cambridge: Harvard UP.

Boyd, Richard. 1988. "How to Be a Moral Realist." in Sayre-McCord, ed. 1988a. pp. 181-228.

Bratman, Michael. 1984. "Two Faces of Intention." in Mele, 1997. pp. 178-203.

Brighouse, M.H. 1990. "Blackburn's Projectivism — An Objection." Philosophical Studies 59 (1990). pp. 225-33.

Brink, David. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge UP.

ibid. 2001. "Realism, Naturalism, and Moral Semantics." Social Philosophy and Policy 18:2. (Summer 2001). pp. 154-76.

Brueckner, Anthony. 1991. "The Omniscient Interpreter Rides Again." Analysis 51:4 (Oct. 1991) pp. 199-205.

Burge, Tyler. 1999. "Comprehension and Understanding." in Hahn, ed. 1999. pp. 229-50.

Burnyeat, M. F. 1992. "Is An Aristotelian Philosophy of Mind Still Credible? (A Draft)." in Nussbaum and Rorty, eds. 1992. pp. 15-26.

Carpenter, Andrew. 1998. "Davidson's Externalism and the Unintelligibility of Massive Error." Disputatio 4: 1998. pp. 25-45.

Carson, Thomas. 1992a. "Gibbard's Conceptual Scheme for Moral Philosophy." Philosophy and Phenomenological Research LII:4 (December 1992). pp. 953-6.

Cooley, D.R. 2006. "Crimina Carnis and Morally Obligatory Suicide." *Ethical Theory and Moral Practice*. 9:3 (June 2006). pp. 327-56.

Cooper, David. 1978. "Moral Relativism." Midwest Studies in Philosophy III (1978). pp. 97-108.

Copp, David. 2000. "Milk, Honey, and the Good Life on Moral Twin Earth." Synthese 124. (2000). pp. 113-37.

Cutrofello, Andrew. 1999. "The Transcendental Pretensions of the Principle of Charity." in Hahn, ed. 1999. pp. 333-41.

Dalmiya, Vrinda. 1990. "Coherence, Truth, and the 'Omniscient Interpreter.'" Philosophical Quarterly 40:158 (Jan 1990). pp. 86-94.

Dancy, Jonathan. 2004. *Ethics Without Principles*. Oxford: Clarendon Press.

Daniels, Norman. 1979. "Wide Reflective Equilibrium and Theory Acceptance in Ethics." Journal of Philosophy 76:5 (May, 1979). pp. 256-82.

Darwall, Stephen. 1983. *Impartial Reason*. Ithaca: Cornell UP.

Davidson, Donald. 1963. "Actions, Reasons, and Causes." in Davidson, 1980. pp. 3-20.

ibid. 1967. "Truth and Meaning." in Davidson, 1984, pp. 17-36.

ibid. 1968. "On Saying That." in Davidson, 1984. pp. 93-108.

ibid. 1970. "Mental Events." in Davidson, 1980. pp. 207-24.

ibid. 1970b. "How is Weakness of the Will Possible?" in Davidson, 1980, pp. 21-43.

ibid. 1973. "The Material Mind." in Davidson, 1980. pp. 245-60.

ibid. 1973b. "Radical Interpretation." in Davidson 1984, pp. 125-40.

ibid. 1974. “On the Very Idea of a Conceptual Scheme.” in Davidson, 1984. pp. 183-198.

ibid. 1977. "Reality Without Reference." in Davidson, 1984, pp. 215-26.

ibid. 1977a. "The Method of Truth in Metaphysics." in Davidson, 1984. pp. 199-214.

ibid. 1978. "Intending." in Davidson, 1980. pp. 83-102.

ibid. 1979. "Moods and Performances." in Davidson, 1984, pp. 109-21.

ibid. 1979b. "The Inscrutability of Reference." in Davidson, 1984, pp. 227-241.

ibid. 1980. *Essays on Actions and Events*. Oxford: Oxford UP.

ibid. 1982. "Rational Animals." in Davidson, 2001. pp. 95-106.

ibid. 1982a. "Empirical Content." in Davidson 2001. pp. 159-76.

ibid. 1983. "A Coherence Theory of Truth and Knowledge." in Davidson, 2001. pp. 137-53.

ibid. 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford UP.

ibid. 1986. "A Nice Derangement of Epitaphs." in Davidson 2005a, pp. 89-108.

ibid. 1988. "Epistemology and Truth." in Davidson, 2001. pp. 177-92.

ibid. 1989. "What is Present to the Mind?" in Davidson, 2001. pp. 53-68.

ibid. 1990. "Epistemology Externalized." in Davidson, 2001. pp. 193-204.

ibid. 1993. "Method and Metaphysics." in Davidson, 2005a. pp. 39-46.

ibid. 1991. "Three Varieties of Knowledge." in Davidson, 2001. pp. 205-20.

ibid. 1995. "The Problem of Objectivity." in Davidson, 2004. pp. 3-18.

ibid. 1996. "The Folly of Trying to Define Truth." in Davidson, 2005a. pp. 19-38.

ibid. 1997. "Truth Rehabilitated." in Davidson, 2005a. pp. 3-18.

ibid. 1997a. “Seeing Through Language.” in Davidson, 2005a, pp. 127-142.

ibid. 1999. "Reply to Neale." in Hahn, 1999. pp. 667-9.

ibid. 1999a. "Reply to A.C. Genova." in Hahn, 1999. pp. 192-4.

ibid. 2001. *Subjective, Intersubjective, Objective*. Oxford: Oxford UP.

ibid. 2001a. "What Thought Requires." in Davidson, 2004. pp. 135-50.

ibid. 2004. *Problems of Rationality*. Oxford: Oxford UP.

ibid. 2005a. *Truth, Language, and History*. Oxford: Oxford UP.

ibid. 2005b. *Truth and Predication*. Cambridge: Belknap Harvard.

De Caro, Mario, ed. 1999. *Interpretations and Causes New Perspectives on Donald Davidson's Philosophy*. Dordrecht: Kluwer Academic Publishers.

De Rosa, Raffaella. 1999. "Is There a Problem About Davidson's Externalism Vis-a-Vis His Holism?" in De Caro, ed., 1999. pp. 201-15.

Divers, John, and Miller, Alexander. 1994. "Why Expressivists about Value Should Not Love Minimalism about Truth." Analysis 54:1 (January 1994). pp. 12-9.

Dummett, Michael. 1959. "Truth." in Dummett, 1978. pp. 1-24.

ibid. 1974a. "The Philosophical Basis of Intuitionistic Logic." in Dummett, 1978. pp. 215-47.

ibid. 1974b. "The Social Character of Meaning." in Dummett, 1978. pp. 420-30.

ibid. 1978. *Truth and Other Enigmas*. Cambridge: Harvard UP.

ibid. 1981. *The Interpretation of Frege's Philosophy*. Cambridge: Harvard UP.

Engel, Pascal. 2001. in Kotatko, Pagin, and Segal, eds. 2001. *Interpreting Davidson*. CSLI Publications. pp. 37-51.

Etchemendy, John. 1988. "Tarski on Truth and Logical Consequence." Journal of Symbolic Logic 53:1 (March 1988). pp. 51-79.

Feyerabend, Paul. 1987. *Farewell to Reason*. New York: Verso.

Fodor, Jerry, and LePore, Ernest. 1992. *Holism A Shopper's Guide*. Cambridge: Blackwell.

Foley, Richard, and Fumerton, Richard. 1985. "Davidson's Theism?" Philosophical Studies 48 (1985) pp. 83-9.

Foot, Philippa. 1958. "Moral Beliefs." in Foot, Philippa. 1978. *Virtues and Vices*. Berkeley: University of California Press. pp. 110-31.

Frege, Gottlob. 1879. *Begriffsscrhift*. selections in Beaney, 1997. pp. 47-78.

ibid. 1884. *The Foundations of Arithmetic*. selections in Beaney, 1997. pp. 84-129.

ibid. 1891. "On Function and Concept." in Beaney, 1997, pp. 130-148.

ibid. 1892a. "On *Sinn* and *Bedeutung*." in Beaney, 1997, pp. 151-71.

ibid. 1892b. "On Concept and Object." in Beaney, 1997, pp. 181-193.

Fuller, Steve. 1988. *Social Epistemology*. Bloomington: Indiana UP.

Geach, P.T. 1960. "Ascriptivism." Philosophical Review 69. pp. 221-5.

ibid. 1965. "Assertion." Philosophical Review 74. pp. 449-65.

Genova, A.C. 1984. "Good Transcendental Arguments." Kant-Studien 75. pp. 469-95.

ibid. 1999. "The Very Idea of Massive Truth." in Hahn, ed. 1999. pp. 167-91.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings A Theory of Normative Judgment*. Cambridge: Harvard UP.

ibid. 1992a. "Précis of *Wise Choices, Apt Feelings*." Philosophy and Phenomenological Research LII:4 (December 1992). pp. 943-5.

ibid. 1992b. "Reply to Blackburn, Carson, Hill, and Railton." Philosophy and Phenomenological Research LII:4 (December 1992). pp. 969-80.

Goldberg, Nathaniel. 2004. "The Principle of Charity." Dialogue XLIII (2004). pp. 671-83.

Grandy, Richard. 1973. "Reference, Meaning, and Belief." Journal of Philosophy 70:14 (August 16, 1973). pp. 439-52

Gupta, Anil. 1993. "A Critique of Deflationism." in Blackburn, Simon, and Simmons, Keith, eds. 1999. *Truth*. Oxford: Oxford UP. pp. 282-307.

Hacking, Ian. 1979. "Foucault's Immature Science." Nous 13:1 (March, 1979). pp. 39-51.

Hahn, Lewis Edwin, ed. 1999. *The Philosophy of Donald Davidson*. Chicago and La Salle: Open Court Publishing.

Haldane, John, and Wright, Crispin, eds. 1993. *Reality, Representation, and Projection*. New York and Oxford: Oxford UP.

Hale, Bob. 1986. "Critical Study: The Compleat Projectivist." The Philosophical Quarterly 36:142 (January 1986). pp. 65-84.

ibid. 1993. "Can There Be a Logic of Attitudes?" in Haldane, John, and Wright, Crispin, eds. 1993. pp. 337-64.

Hanson, Norwood Russell. 1958. *Patterns of Discovery*. Cambridge: Cambridge UP.

Hare, R.M. 1952. *The Language of Morals*. Oxford: Clarendon Press.

Harman, Gilbert. 1977. *The Nature of Morality An Introduction to Ethics*. Oxford: Oxford UP.

ibid. 1986. "Moral Explanations of Natural Facts— Can Moral Claims be Tested Against Moral Reality?" Southern Journal of Philosophy XXIV (1986) supplement. pp. 57-67.

Hayek, Friedrich. 1952. *The Sensory Order*. Chicago: University of Chicago Press.

ibid. 1955. "Degrees of Explanation." in Hayek, 1967. pp. 3-21.

ibid. 1963. "Rules, Perception, and Intelligibility." in Hayek, 1967. pp. 43-65.

ibid. 1964. "The Theory of Complex Phenomena." in Hayek, 1967. pp. 22-42.

ibid. 1967. *Studies in Philosophy, Politics, and Economics*. Chicago: University of Chicago Press.

Hill, Thomas. 1992. "Gibbard on Morality and Sentiment." Philosophy and Phenomenological Research LII:4 (December 1992). pp. 957-60.

Hintikka, Jaakko. 1969a. "Deontic Logic and its Philosophical Morals." in Hintikka, Jaakko. 1969b. *Models for Modalities Selected Essays*. Dordrecht: D. Reidel Publishing Company. pp. 184-214.

Hochberg, Herbert. 2003. *Introducing Analytic Philosophy Its Sense and its Nonsense 1879-2002*. Frankfurt: Ontos Verlag.

Holton, Richard. 2002. "Principles and Particularisms." Proceedings of the Aristotelian Society Supplementary Volume 76. pp. 191-210.

Hooker, Brad, and Little, Margaret, eds. 2000. *Moral Particularism*. Oxford: Oxford UP.

Horgan, Terence, and Timmons, Mark. 1990. "New Wave Moral Realism Meets Moral Twin Earth." Journal of Philosophical Research 16 (1990-1). pp. 447-65.

ibid. 1992. "Troubles on Moral Twin Earth: Moral Queerness Revisited." Synthese 92. (1992). pp. 221-60.

Horwich, Paul. 1990/1998. *Truth 2nd ed.* Oxford: Oxford UP.

ibid. 1993. "Gibbard's Theory of Norms." Philosophy and Public Affairs 22:1 (Winter 1993). pp. 67-78.

ibid. 1994. "The Essence of Expressivism." Analysis 54:1 (January 1994). pp. 19-20.

ibid. "Meaning, Use, and Truth." in Horwich, Paul. 2004. *From a Deflationary Point of View*. Oxford: Oxford UP. pp. 67-85.

ibid. 1998. *Meaning*. Oxford: Oxford UP.

ibid. 1999. "Davidson on Deflationism." in Zeglen, Urszula, ed. 1999. pp. 20-24.

Hurley, Susan. 1985. "Objectivity and Disagreement." in Honderich, Ted. ed. 1985. *Morality and Objectivity*. London: Routledge and Kegan Paul. pp. 54-97.

Jackson, Frank, Oppy, Graham, and Smith, Michael. 1994. "Minimalism and Truth-Aptness." Mind 103:411 (July 1994). pp. 287-302.

Jackson, Frank; Pettit, Philip; and Smith, Michael. 2000. "Ethical Particularism and Patterns." in Hooker and Little, eds. 2000. pp. 79-99.

Kant, Immanuel. 1787/1996. *Critique of Pure Reason*. trans. Pluhar, Werner. Indianapolis: Hackett Publishing Company.

ibid. 1785/1949. *Fundamental Principles of the Metaphysics of Morals*. trans. Abbott, Thomas. New York: Macmillan/Library of Liberal Arts.

ibid. 1788/1993. *Critique of Practical Reason*. trans. Beck, Lewis White. Upper Saddle River: Prentice Hall.

ibid. 1997. *Lectures on Ethics*. eds. Heath, Peter and Schneewing, J. B.; trans. Heath, Peter. Cambridge: Cambridge UP.

Kim, Jaegwon, 1984. "Concepts of Supervenience." in Kim, Jaegwon. 1993. *Supervenience and Mind Selected Philosophical Essays*. Cambridge: Cambridge UP. pp. 53-78.

Kirkham, Richard. 1992. *Theories of Truth A Critical Introduction*. Cambridge: MIT Press.

Klein, Peter. 1986. "Radical Interpretation and Global Skepticism." in LePore, Ernest, ed. 1986. pp. 369-86.

Korsgaard, Christine. 1986. "Skepticism About Practical Reason." in Korsgaard, 1996. *Creating the Kingdom of Ends*. Cambridge: Cambridge UP.

ibid. 1996a. *The Sources of Normativity*. with Cohen, G.A.; Geuss, Raymond; Nagel, Thomas; and Williams, Bernard. O'Neill, Onora ed. Cambridge: Cambridge UP.

Kripke, Saul. 1972. *Naming and Necessity*. Cambridge: Harvard UP.

ibid. 1979. "A Puzzle About Belief." in Salmon, Nathan and Soames, Scott. eds. 1988. *Propositions and Attitudes*. Oxford: Oxford UP. pp. 102-48.

Kuhn, Thomas. 1962/1970. *The Structure of Scientific Revolutions 2nd edition, enlarged*. Chicago: Chicago UP.

Leibniz, Gottfried Wilhelm von. 1686/1902. *Discourse on Metaphysics, Correspondence with Arnauld, Monadology*. La Salle: Open Court.

Lemos, John. 2000. "The Problems With Emotivism: Reflections on Some MacIntyrean Arguments." Journal of Philosophical Research XXV (2000). 285-309.

Lepore, Ernest. ed. 1986. *Truth and Interpretation*. New York: Blackwell.

LePore, Ernest. 1999. "Davidson and Understanding Language." in De Caro, ed. 1999. pp. 47-70.

Lepore, Ernie, and Ludwig, Kirk. 2005. *Davidson: Meaning, Truth, Language, and Reality*. Oxford: Clarendon Press.

Lewin, Ronald. 1978. *Ultra Goes to War*. New York: McGraw-Hill.

Lewis, David. 1974. "Radical Interpretation." Synthese 23 (1974). pp. 331-44.

Little, Margaret. 1997. "Virtue as Knowledge: Objections from the Philosophy of Mind." Nous 31:1 (March, 1997). pp. 59-79.

ibid. 2000. "Moral Generalities Revisited." in Hooker and Little, eds., 2000. pp. 276-304.

Ludwig, Kirk. 1992. "Skepticism and Interpretation." Philosophy and Phenomenological Research 52:2 (June 1992). pp. 317-39.

Lynch, Michael. 2004. "Minimalism and the Value of Truth." Philosophical Quarterly 54:217 (October 2004). pp. 497-517.

MacIntyre, Alasdair. 1984. *After Virtue A Study in Moral Theory*. *2nd ed.* Notre Dame: University of Notre Dame Press.

Mackie, J. L. 1977. *Ethics Inventing Right and Wrong*. New York: Penguin.

Maker, William. 1991. "Davidson's Transcendental Arguments." Philosophy and Phenomenological Research 51:2 (June 1991). pp. 345-60.

Marton, Peter. 1999. "Omniscient versus Super-Omniscient Interpreters." Philosophical Quarterly 49:194 (Jan 1999). pp. 72-7.

McDowell, John. 1978. "Are Moral Requirements Hypothetical Imperatives?" in McDowell, 1998. pp. 77-94.

ibid. 1981. "Non-Cognitivism and Rule-Following." in McDowell, 1998. pp. 198-220.

ibid. 1985. "Values and Secondary Qualities." in McDowell, 1998. pp. 131-50.

ibid. 1998. *Mind, Value, and Reality*. Cambridge: Harvard UP.

McGinn, Colin. 1977. "Charity, Interpretation, and Belief." Journal of Philosophy 74:9 (September 1977). pp. 521-35.

ibid. 1986. "Radical Interpretation and Epistemology." in LePore, ed. 1986. pp. 356-68.

McNaughton, David, and Rawling, Piers. 2000. "Unprincipled Ethics." in Hooker and Little, 2000, eds. pp. 256-75.

Mele, Alfred, ed. 1997. *The Philosophy of Action*. Oxford: Oxford UP.

Merli, David. 2002. "Return to Moral Twin Earth." Canadian Journal of Philosophy 32:2. (June, 2002). pp. 207-40.

Mises, Ludwig von. 1949/1966. *Human Action A Treatise on Economics Third Revised Edition*. Chicago: Contemporary Books.

Moore, G. E. 1903/1993. *Principia Ethica*. Revised Ed., ed. Baldwin, Thomas. Cambridge: Cambridge UP.

Mulligan, Kevin; Simons, Keith; and Smith, Barry. 1984. "Truth-Makers." Philosophy and Phenomenological Research. 44:3. (March, 1984). pp. 287-321.

Nagel, Thomas. 1970. *The Possibility of Altruism*. Oxford: Clarendon Press.

Neale, Stephen. 2001. *Facing Facts*. Oxford: Oxford UP.

Norris, Christopher. 1985. *The Contest of Faculties Philosophy and Theory after Deconstruction*. London and New York: Methuen.

Nussbaum, Martha. 1978. *Aristotle's* De Motu Animalium. Princeton: Princeton UP.

Nussbaum, Martha, and Putnam, Hilary. 1992. "Changing Aristotle's Mind." in Nussbaum and Rorty, eds. 1992. pp. 27-56.

Nussbaum, Martha, and Rorty, Amélie Oksenberg, eds. 1992. *Essays on Aristotle's* De Anima. Oxford: Clarendon Press.

Ogden, C.K., and Richards, I.A. 1923/1956. *The Meaning of Meaning A Study of the Influence of Language upon Thought and of the Science of Symbolism*. London: Routledge and Kegan Paul.

Price, Huw. 1994. "Semantic Minimalism and the Frege Point." in Tsohatzidis, Savas, ed. 1994. *Foundations of Speech Act Theory Philosophical and Linguistic Perspectives*. London and New York: Routledge. pp. 132-55.

Putnam, Hilary. 1960. "Minds and Machines." in Putnam, 1975. pp. 362-85.

ibid. 1973. "Philosophy and Our Mental Life." in Putnam, 1975. pp. 291-303.

ibid. 1975. *Mind, Language, and Reality Philosophical Papers Volume 2*. Cambridge: Cambridge UP.

ibid. 1975a. "The Meaning of 'Meaning'." in Putnam 1975. pp. 215-71.

ibid. 1991. "Does the Disquotational Theory of Truth Solve All Philosophical Problems?" in Putnam, Hilary. 1994. *Words and Life*. Cambridge: Harvard UP. pp. 264-278.

Punzo, Vincent. 1969. *Reflective Naturalism*. New York: Macmillan.

Quine, Willard Van Orman. 1960. *Word and Object*. Cambridge: MIT Press.

Railton, Peter. 1992. "Nonfactualism about Normative Discourse." Philosophy and Phenomenological Research LII:4 (December 1992). pp. 961-8.

Ramberg, Bjorn. 1989. *Donald Davidson's Philosophy of Language An Introduction*. Oxford: Basil Blackwell.

Rawls, John. 1971/1999. *A Theory of Justice Revised Edition.* Cambridge: Harvard UP.

ibid. 1975. "The Independence of Moral Theory." in Rawls, John. 1999. *Collected Papers*. ed. Freeman, Samuel. Cambridge: Harvard UP. pp. 286-302.

ibid. 1993/1996. *Political Liberalism with a New Introduction and the "Reply to Habermas*." New York: Columbia UP.

ibid. 2000. *Lectures on the History of Moral Philosophy*. ed. Herman, Barbara. Cambridge: Harvard UP.

Register, Bryan. 2003. *The Critique of Political Reason in Political Reason in Friedrich Hayek and Michel Foucault*. Unpublished Master's Report.

Ross, Peter. 2002. "Explaining Motivated Desires." Topoi 21. pp. 199-207.

Russell, Bertrand. 1905. "On Denoting." in Marsh, Robert, ed. 1956. *Logic and Knowledge*. London: Routledge. pp. 39-56.

ibid. 1913/1984. *Theory of Knowledge The 1913 Manuscript*. London: Routledge and Kegan Paul.

Sayre-McCord, Geoffrey, ed. 1988a. *Essays on Moral Realism*. Ithaca: Cornell UP.

ibid. 1988b "Introduction: The Many Moral Realisms." in Sayre-McCord, ed., 1988a. pp. 1-23.

Scanlon, T.M. 1998. *What We Owe to Each Other*. Cambridge: Harvard UP.

Schiffer, Stephen. 1987. *Remnants of Meaning*. Cambridge: MIT Press.

Searle, John. 1962. "Meaning and Speech Acts." Philosophical Review 71. pp. 423-32.

ibid. 1969. *Speech Acts An Essay in the Philosophy of Language*. Cambridge: Cambridge UP.

ibid. 1979. *Expression and Meaning Studies in the Theory of Speech Acts*. Cambridge: Cambridge UP.

ibid. 1983. *Intentionality*. Cambridge: Cambridge UP.

ibid. 1987. "Indeterminacy, Empiricism, and the Third Person." in Searle, John. 2002. *Consciousness and Language*. Cambridge: Cambridge UP. pp. 226-50.

Searle, John, and Vanderveeken, Daniel. 1985. *Foundations of Illocutionary Logic*. Cambridge: Cambridge UP.

Sellars, Wilfrid. 1956/1997. *Empiricism and the Philosophy of Mind*. Cambridge: Harvard UP.

Schueler, G.F. 1988. "Modus Ponens and Moral Realism." Ethics 98 (April 1988). pp. 492-500.

Smith, Barry. 1994. *Austrian Philosophy The Legacy of Franz Brentano*. Chicago: Open Court Publishing.

ibid. 1998. "On Knowing One's Own Language." in Wright, Smith, and Macdonald, eds. 1998. pp. 391-428.

Smith, Michael. 1994. "Why Expressivists about Value should Love Minimalism about Truth." Analysis 54:1 (January 1994). pp. 1-12.

ibid. 1994a. "Minimalism, Truth-Aptitude, and Belief." Analysis 54:1 (January 1994). pp. 21-6.

ibid. 1994b. *The Moral Problem*. Malden: Blackwell Publishing.

Soames, Scott. 1999. *Understanding Truth*. New York: Oxford UP.

Soble, Alan. 2003. "Kant and Sexual Perversion." Monist 86:1 (Jan 2003). pp. 55-89.

Sosa, David. 1996. "The Import of the Puzzle About Belief." Philosophical Review 105:3 (July, 1996). pp. 373-402.

Stevenson, Charles. 1944. *Ethics and Language*. New Haven and London: Yale UP.

ibid. 1963. *Facts and Values Studies in Ethical Analysis*. New Haven & London: Yale UP.

Stocker, Michael. 1979. "Desiring the Bad: An Essay in Moral Psychology." Journal of Philosophy 76:12 (Dec. 1979). pp. 738-53.

Stoljar, Daniel. 1992. "Emotivism and Truth Conditions." Philosophical Studies 70 (1993). pp. 81-101.

Strawson, P.F. 1966. *The Bounds of Sense An Essay on Kant's Critique of Pure Reason*. New York: Routledge.

Stroud, Barry. 1999. "Radical Interpretation and Philosophical Skepticism." in Hahn, 1999. pp. 139-61.

Sturgeon, Nicholas. 1984. "Moral Explanations." Copp, David, and Zimmerman, David, eds. 1984. *Morality, Reason, and Truth*. Totowa: Rowman & Allanheld. pp. 49-78.

ibid. 1986a. "Harman on Moral Explanations of Natural Facts." Southern Journal of Philosophy XXIV (1986) supplement. pp. 69-78.

ibid. 1986b. "What Difference Does it Make Whether Moral Realism is True?" Southern Journal of Philosophy XXIV (1986) supplement. pp. 115-41.

ibid. 1990. "Contents and Causes: A Reply to Blackburn." Philosophical Studies 61 (1991). pp. 19-37.

Tarski, Alfred. 1933. "The Concept of Truth in Formalized Languages." in Tarski, 1983. pp. 152-278.

ibid. 1936. "The Establishment of Scientific Semantics." in Tarski, 1983. pp. 401-8.

ibid. 1944. "The Semantic Conception of Truth." in Linsky, Leonard, ed. 1952/1970. *Semantics and the Philosophy of Language*. Urbana: University of Illinois Press. pp. 13-49.

ibid. 1983. *Logic, Semantics, Metamathematics. 2nd ed*. ed, Corcoran, John. Indianapolis: Hackett Publishing Company.

Tennant, Neil. 1999. "Radical Interpretation, Logic, and Conceptual Schemes." in de Caro, ed., 1999. pp. 71-94.

Tersman, Folke. 1995. "Non-Cognitivism and Inconsistency." Southern Journal of Philosophy XXXIII:3 (1995). pp. 361-71.

Twain, Mark. 1880. *A Tramp Abroad*. in Twain, Mark. 1992. *The Family Mark Twain*. New York: Barnes and Noble Books.

Unwin, Nicholas. 1990. "Can Emotivism Sustain a Social Ethics?" Ratio III:1 (June 1990). pp. 64-81.

Urmson, J. O. 1968. *The Emotive Theory of Ethics*. London: Hutchinson University Library.

Vahid, Hamid. 2001. "Charity, Supervenience, and Skepticism." Metaphilosophy 32:3 (April 2001). pp. 308-25.

van Roojen, Mark. 1996. "Expressivism and Irrationality." Philosophical Review 105:3 (July 1996). pp. 311-35.

Velleman, J. David. 1996. "The Possibility of Practical Reason." in Velleman, J. David. 2000. *The Possibility of Practical Reason*. Oxford: Clarendon Press. pp. 170-99.

Vermazen, Bruce. 1986. "Testing Theories of Interpretation." in LePore, ed. 1986. pp. 235-244.

Wallace, John. 1986. "Translation Theories and the Decipherment of Linear B." in LePore, ed. 1986. pp. 211-34.

Wallace, R. Jay. 1990. "How to Argue about Practical Reason." in Mind 99:395 (July 1990). pp. 355-85.

Wedgwood, Ralph. 1997. "Non-Cognitivism, Truth, and Logic." Philosophical Studies 86 (1997). pp. 73-91.

Wilks, Colin. 2002. *Emotion, Truth, and Meaning In Defense of Ayer and Stevenson*. Dordrecht, Boston, London: Kluwer Academic Publishers.

Wiggins, David. 1976. "Truth, Invention, and the Meaning of Life." in Wiggins, 1987. *Needs, Values, Truth*. Oxford: Blackwell. pp. 87-138.

Williams, Bernard. 1974. "The Truth in Relativism." in Williams, 1981. pp. 132-43.

ibid. 1980. "Internal and External Reasons." in Williams, 1981. pp. 101-13.

ibid. 1981. *Moral Luck*. Cambridge: Cambridge UP.

ibid. 1985. *Ethics and the Limits of Philosophy*. Cambridge: Harvard UP.

Wittgenstin, Ludwig. 1922. *Tractatus Logico-Philosophicus*. London: Routledge. trans Ogden, C.K.

Wright, Crispin; Smith, Barry; and Macdonald, eds. 1998. *Knowing Our Own Minds*. Oxford: Clarendon Press.

Zangwill, Nick. 1992. "Moral Modus Ponens." Ratio V:2 (December 1992). pp. 177-93.

Zeglen, Urszula, ed. 1999. *Donald Davidson Truth, Meaning and Knowledge*. London: Routledge.

Vita

Bryan Randall Register was born January 22, 1976, in Raleigh, North Carolina. His parents are David Sheppard Register and Bridget Mintz Testa. He graduated from Clear Lake High School in Houston, Texas, in 1994, and entered the College of Liberal Arts of the University of Texas at Austin. He married Sarah Jo French on August 1, 1998. He took the degrees of Bachelor of Arts in Philosophy and History and Bachelor of Science in Speech Communication in 1999. Remaining at the University, he took the degree of Master of Arts in Philosophy in 2003, for his report on Friedrich Hayek and Michel Foucault. His daughter, Eilonwy Kenna Register, was born June 7, 2006.

Permanent address: 1213 Meadgreen Dr., Austin, TX 78758

This dissertation was typed by the author.