Papers in History and Methodology of Analytic Philosophy

Brian Weatherson

March 5, 2014

Contents

What Good are Counterexamples?

The following kind of scenario is familiar throughout analytic philosophy. A bold philosopher proposes that all *F*s are *G*s. Another philosopher proposes a particular case that is, intuitively, an *F* but not a *G*. If intuition is right, then the bold philosopher is mistaken. Alternatively, if the bold philosopher is right, then intuition is mistaken, and we have learned something from philosophy. Can this alternative ever be realised, and if so, is there a way to tell when it is? In this paper, I will argue that the answer to the first question is *yes*, and that recognising the right answer to the second question should lead to a change in some of our philosophical practices.

The problem is pressing because there is no agreement across the sub-disciplines of philosophy about what to do when theory and intuition clash. In epistemology, particularly in the theory of knowledge, and in parts of metaphysics, particularly in the theory of causation, it is almost universally assumed that intuition trumps theory. Shope's *The Analysis of Knowledge* contains literally dozens of cases where an interesting account of knowledge was jettisoned because it clashed with intuition about a particular case. In the literature on knowledge and lotteries it is not as widely assumed that intuitions about cases are inevitably correct, but this still seems to be the working hypothesis.[1] And recent work of causation by a variety of authors, with a wide variety of opinions, generally takes the same line: if a theory disagrees with intuition about a case, the theory is wrong.[2] In this area exceptions to the rule are a little more frequent, particularly on the issues of whether causation is transitive and whether omissions can be causes, but in most cases the intuitions are taken to override the theories. Matters are quite different in ethics. It is certainly not a good thing for utilitarian theories that we very often feel that the action that maximises utility is *not* the right thing to do. But the existence of such cases is rarely taken to be obviously and immediately fatal for utilitarian theories in the way that, say, Gettier cases are taken to be obviously and immediately fatal for theories of knowledge that proclaim those cases to be cases of knowledge. Either there is some important difference here between the anti-utilitarian cases and the Gettier cases, a difference that justifies our differing reactions, or someone is making a mistake. I claim that it is (usually) the epistemologists and the metaphysicians who are wrong. In more cases than we usually imagine, a good philosophical theory can teach us that our intuitions are mistaken. Indeed, I think it is possible (although perhaps not likely) that the justified true belief (hereafter, JTB) theory of knowledge is so plausible that we should hold onto it in preference to keeping our intuition that Gettier cases are not cases of knowledge.

My main interests here are methodological, not epistemological. Until the last section I will be arguing for the JTB theory of knowledge, but my main interest is in showing that one particular argument against the JTB theory, the one that turns on the fact that it issues in some rather unintuitive pronouncements about Gettier cases, is not in itself decisive. Still, the epistemological issues are important, which is one reason I chose to focus on the JTB theory,

† Penultimate draft only. Please cite published version if possible. Final version published in *Philosophical Studies* 115 (2003): 1-31.

[1] See, for example, DeRose (1996) and Nelkin (2000)

[2] See, for example, Menzies (1996), or any of the papers in the special *Journal of Philosophy* issue on causation, April 2000.

and at the end I will discuss how the methodological conclusions drawn here may impact on them in an unexpected way.

1 Intuitions

Let us say that a **counterexample** to the theory that all *F*s are *G*s is a possible situation such that most people have an intuition that some particular thing in the story is an *F* but not a *G*. The kinds of intuition I have in mind are what George Bealer (1998) calls intellectual "seemings". Bealer distinguishes intellectual seemings, such as the intuition that Hume's Principle is true, or that punishing a person for a crime they did not commit is unjust, from physical seemings, such as the 'intuition' that objects fall if released, or perhaps that the sun rotates around the earth. We shall be primarily concerned here with intellectual seemings, and indeed I shall only call these intuitions in what follows.

As Bealer notes, whether something seems to be true can be independent of whether we believe it to be true. Bealer himself notes that Frege's Axiom V seems to be true, though we know it is false. It does not seem to be the case, in the relevant sense, that 643 x 721 = 463603. Unless one is rather good at mental arithmetic, there is nothing that 643 x 721 seems to be; it is out of the reach of intuition. These are not the only ways that seemings and belief can come apart. One can judge that something seems to be the case while neither believing nor disbelieving it. This is a sensible attitude to take towards the view that one cannot *know* that a particular ticket will lose in a fair lottery. This is despite the fact that it certainly *seems* one cannot know this. If one's intuitions are running rampant, one may even have an intuition about something that one believes to be strictly indeterminate. For example, some people may have the intuition that the continuum hypothesis is true, even though they believe on reflection that it is indeterminate whether it is true.

The distinction between intuitions and belief is important because it helps reduce the violence that revisionary philosophical views do to our pre-existing positions. When I say that Gettier cases may be cases of knowledge, I am not denying that there is a strong intuition that they are not cases of knowledge. I am not denying that a Gettier case does not *seem* to be a case of knowledge. The same thing occurs in ethics. Utilitarians rarely deny that it seems that punishing innocents is the wrong thing to do. They urge that in certain, rare, cases this might be one of those things that seems to be true despite being false. The case that knowledge is justified true belief is meant to be made in full awareness of the fact that certain cases of justified true beliefs seem to not be cases of knowledge.

Actually, although we will not make much of it here, this last claim is not true as a general statement about all people. Jonathan Weinberg, Stephen Stich and Shaun Nichols have reported Weinberg et al. (2001) that the intuition that Gettier cases are not cases of knowledge is not universally shared. It is not entirely clear what the philosophical relevance of these discoveries is. It *might* show that we who have Gettier intuitions speak a different language from those who do not. It *might* show (though as Stich and Nichols point out it is rather hard to see how) that philosophers know a lot more about knowledge than other folk. I think it is rather unlikely that this is true, but we shall bracket such concerns for now, and continue on the assumption that all parties have the Gettier intuitions. Since I shall want to argue that knowledge may still be justified belief in any case, I am hardly tilting the playing field in my direction by making this assumption.

Given that intuitions are what Bealer calls intellectual seemings, and given that the example of Axiom V shows that seemings can be mistaken, what evidence have we that they are not mistaken in the cases we consider here? Arguably, we have very little indeed. Robert Cummins (1998) argues that in general intuition should not be trusted as an evidential source because it cannot be calibrated. We wouldn't have trusted the evidence Galileo's telescope gave us about the moon without an independent reason for thinking his telescope reliable. Fortunately, this can be done; we can point the telescope at far away terrestrial mountains, and compare its findings with the findings of examining the mountains up close and personal. There is no comparable way of calibrating intuitions. Clearly we should suspicious of any method that has been tested and found unreliable, but there are tricky questions about the appropriate level of trust in methods that have not been tested. Ernest Sosa (1998) argues in response to Cummins that this kind of reasoning leads to an untenable kind of scepticism. Sosa notes that one can make the same point about perception as Cummins makes about intuition: we have no independent way of calibrating perception as a whole. There is a distinction to be drawn here, since perception divides into natural kinds, visual perception, tactile perception, etc, and we can use each of these to calibrate the others. It is hard to see how intuitions can be so divided in ways that permit us to check some kinds of intuitions against the others. In any case, the situation is probably worse than Cummins suggests, since we know that several intuitions are just false. It is interesting to note the many ways in which intuition does, by broad agreement, go wrong.

Many people are prone to many kinds of systematic **logical** mistakes. Most famously, the error rates on the Wason Selection Task are disturbingly large. Although this test directly measures beliefs rather than intuitions, it seems very likely that many of the false beliefs are generated by mistaken intuitions. As has been shown in a variety of experiments, the most famous of which were conducted by Kahneman and Tversky, most people are quite incompetent at **probabilistic** reasoning. In the worst cases, subjects held that a conjunction was more probable than one of its conjuncts. Again, this only directly implicates subjects' beliefs, but it is very likely that the false beliefs are grounded in false intuitions. (The examples in this paragraph are discussed in detail in Stich (1988, 1992).)

As noted above, most philosophers would agree that many, if not most, people have mistaken **moral** intuitions. We need not agree with those consequentialists who think that vast swathes of our moral views are in error to think that (a) people make systematic moral mistakes and (b) some of these mistakes can be traced to mistaken intuitions. To take the most dramatic example, for thousands of years it seemed to many people that slavery was morally acceptable. On a more mundane level, many of us find that our intuitive judgements about a variety of cases cannot be all acceptable, for it is impossible to find a plausible theory that covers them all.[3] Whenever we make a judgement inconsistent with such an intuition, we are agreeing that some of our original intuitions were mistaken.

From a rather different direction, there are many mistaken **conceptual** intuitions, with the error traceable to the way Gricean considerations are internalised in the process of learning a language. Having learned that it would be improper to use t to describe a particular case, we can develop the intuition that this case is not an F, where F is the property denoted by t. For

[3]The myriad examples in Unger (1996) are rather useful for reminding us just how unreliable our moral intuitions are, and how necessary it is to employ reflection and considered judgement in regimenting such intuitions.

example, if one is careless, one can find oneself sharing the intuition expressed by Ryle in *The Concept of Mind* that morally neutral actions, like scratching one's head, are neither voluntary nor involuntary (Ryle, 1949). The source of this intuition is the simple fact that it would be odd to describe an action as voluntary or involuntary unless there was some reason to do so, with the most likely such reason being that the action was in some way morally suspect. The fact that the intuition has a natural explanation does not stop it being plainly false. We can get errors in conceptual intuitions from another source. At one stage it was thought that whales are fish, that the Mars is a star, the sun isn't. These are beliefs, not intuitions, but there are clearly related intuitions. Anyone who had these beliefs would have had the intuition that in a situation like *this* (here demonstrating the world) the object in the Mars position was a star, and the objects in the whale position were fish. The empirical errors in the person's belief will correlate to conceptual errors in their intuition. To note further that the kind of error being made here is conceptual not empirical, and hence the kind of error that occurs in intuition, note that we need not have learned anything new about whales, the sun or Mars to come to our modern beliefs. (In fact we did, but that's a different matter.) Rather, we need only have learned something about the vast bulk of the objects that are fish, or stars, to realise that these objects had been wrongly categorised. The factor we had thought to be the most salient similarity to the cases grouped under the term, being a heavenly body visible in the night sky for 'star', living in water for 'fish', turned out not to be the most important similarity between most things grouped under that term. So there is an important sense in which saying whales are fish, or that the sun is not a star, may reveal a conceptual (rather than an empirical) error.

There seems to be a link between these two kinds of conceptual error. The reason we say that the Rylean intuitions, or more generally the intuitions of what (Grice, 1989, Ch. 1) called the Type-A philosophers, are mistaken is that the rival, Gricean, theory attaches to each word a relatively natural property. There is no natural property that actions satisfy when, and only when, we ordinarily describe them as voluntary. There is a natural property that covers all these cases, and other more mundane actions like scratching one's head, and that is the property we now think is denoted by 'voluntary'. This notion of naturalness, and the associated drive for systematicity in our philosophical and semantic theories, will play an important role in what follows.

2 Correcting Mistakes

The following would be a bad defence of the JTB theory against counterexamples. We can tell that all counterexamples to the JTB theory are based on mistaken intuitions, because the JTB theory is true, so all counterexamples to it are false. Unless we have some support for the crucial premise that the JTB theory is true, this argument is rather weak. And that support should be enough to not only make the theory *prima facie* plausible, but so convincing that we are prepared to trust it rather than our judgements about Gettier cases.

In short, the true theory of knowledge is the one that does best at (a) accounting for as many as possible of our intuitions about knowledge while (b) remaining systematic. A 'theory' that simply lists our intuitions is no theory at all, so condition (b) is vital. And it is condition (b), when fully expressed, that will do most of the work in justifying the preservation of the JTB theory in the face of the counterexamples.

The idea that our theory should be systematic is accepted across a wide range of philosophical disciplines. This idea seems to be behind the following plausible claims by Michael Smith: "Not only is it a platitude that rightness is a property that we can discover to be instantiated by engaging in rational argument, it is also a platitude that such arguments have a characteristic coherentist form." (1994: 40) The second so-called platitude just points out that it is a standard way of arguing in ethics to say, you think we should do X in circumstances C_1, circumstances C_2 are just like C_1, so we should do X in C_1. The first points out that not only is this standard, it can yield surprising ethical knowledge. But this is only plausible if it is more important that final ethics is systematic than that first ethics, the ethical view delivered by intuition, is correct. In other words, it is only plausible if ethical intuitions are classified as mistaken to the extent that they conflict with the most systematic plausible theory. So, for example, it would be good news for utilitarianism if there was no plausible rival with any reasonable degree of systematicity.

This idea also seems to do important work in logic. If we just listed intuitions about entailment, we would have a theory on which disjunctive syllogism (A and $\neg A \vee B$ entail B) is valid, while *ex falso quadlibet* (A and $\neg A$ entail B) is not. Such a theory is unsystematic because no concept of entailment that satisfies these two intuitions will satisfy a generalised transitivity requirement: that if C and D entail E, and F entails D then C and F entail E. (This last step assumes that $\neg A$ entails $\neg A \vee B$, but that is rarely denied.) Now one can claim that a theory of entailment that gives up this kind of transitivity can still be systematic enough, and Neil Tennant (1992) does exactly this, but it is clear that we have a serious cost of the theory here, and many people think avoiding this cost is more important than preserving all intuitions.

In more detail, there are four criteria by which we can judge a philosophical theory. First, counterexamples to a theory count against it. While a theory can be reformist, it cannot be revolutionary. A theory that disagreed with virtually all intuitions about possible cases is, for that reason, false. The theory: *X knows that p iff X exists and p is true* is systematic, but hardly plausible. As a corollary, while intuitions about any particular possible case can be mistaken, not too many of them could be. Counterexamples are problematic for a theory, the fewer reforms needed the better, it's just not that they are not fatal. Importantly, not all counterexamples are as damaging to a theory as others. Intuitions come in various degrees of strength, and theories that violate weaker intuitions are not as badly off as those that violate stronger intuitions. Many people accept that the more obscure or fantastic a counterexample is, the less damaging it is to a theory. This seems to be behind the occasional claim that certain cases are "spoils to the victor" – the idea is that the case is so obscure or fantastic that we should let theory rather than intuition be our guide. Finally, if we can explain why we have the mistaken intuition, that counts for a lot in reducing the damage the counterexample does. Grice did not just assert that the theory on which an ordinary head scratch was voluntary was more systematic than the theory of voluntariness Ryle proposed, he provided an explanation of why it might seem that his theory was wrong in certain cases.

Secondly, the analyses must not have too many *theoretical* consequences which are unacceptable. Consider Kahneman and Tversky's account of how agents actually make decisions, prospect theory, as an analysis of 'good decision'. (Disclaimer: This is not how Kahneman and Tversky intend it.) So the analysis of 'good decision' is 'decision authorised by prospect theory'. It is a consequence of prospect theory that which decision is "best" depends on which

outcome is considered to be the neutral point. In practice this is determined by contextual factors. Redescribing a story to make different points neutral, which can be done by changing the context, licences different decisions. I take it this would be unacceptable in an analysis of 'good decision', even though it means the theory gives intuitively correct results in *more* possible cases than its Bayesian rivals[4]. In general, we want our normative theories to eliminate arbitrariness as much as possible, and this is usually taken to be more important than agreeing with our pre-theoretic intuitions about particular cases. Unger uses a similar argument in *Living High and Letting Die* to argue against the reliance on intuitions about particular cases in ethics. We have differing ethical intuitions towards particular cases that differ only in the conspicuousness of the suffering caused (or not prevented), we know that conspicuousness is not a morally salient difference, so we should stop trusting the particular intuitions. (Presumably this is part of the reason that we find Tennant's theory of entailment so incredible, *prima facie*. It is not just that violating transitivity seems unsystematic, it is that we have a theoretical intuition that transitivity should be maintained.)

Thirdly, the concept so analysed should be theoretically significant, and should be analysed in other theoretically significant terms. This is why we now analyse 'fish' in such a way that whales aren't fish, and 'star' in such a way that the sun is a star. This is not just an empirical fact about our language. Adopting such a constraint on categories is a precondition of building a serious classificatory scheme, so it is a constraint on languages, which are classificatory schemes *par excellance*. Even if I'm wrong about this, the fact that we do reform our language with the advance of science to make our predicates refer to theoretically more significant properties shows that we have a commitment to this restriction.

Finally, the analysis must be simple. This is an important part of why we don't accept Ryle's analysis of 'voluntary'. His analysis can explain all the intuitive data, even without recourse to Gricean implicature, and arguably it doesn't do *much worse* than the Gricean explanation on the second and third tests. But Grice's theory can explain away the intuitions that it violates, and importantly it does so merely with the aid of theories of pragmatics that should be accepted for independent reasons, and it is simpler, so it trumps Ryle's theory.

My main claim is that even once we have accepted that the JTB theory seems to say the wrong thing about Gettier cases, we should still keep an open mind to the question of whether it is true. The right theory of knowledge, the one that attributes the correct meaning to the word 'knows', will do best on balance at these four tests. Granted that the JTB theory does badly on test one, it seems to do better than its rivals on tests two, three and four, and this may be enough to make it correct.

3 Naturalness in a theory of meaning

Let's say I have convinced you that it would be better to use 'knows' in such a way that we all now assent to "She knows" whenever the subject of that pronoun truly, justifiably, believes. You may have been convinced that only by doing this will our term pick out a natural relation, and there is evident utility in having our words pick out relations that carve nature at something like its joints. Only in that way, you may concede, will our language be a decent classificatory scheme of the kind described above, and it is a very good thing to have one's language be a decent

[4]A point very similar to this is made in Horowitz (1998).

classificatory scheme. I have implicitly claimed above that if you concede this you should agree that I will have thereby corrected a *mistake* in your usage. But, an objector may argue, it is much more plausible to say that in doing so I simply changed the meaning of 'knows' and its cognates in your idiolect. The meaning of your words is constituted by your responses to cases like Gettier cases, so when I convince you to change your response, I change the meaning of your words.

This objection relies on a faulty theory of meaning, one that equates meaning with use in a way which is quite implausible. If this objection were right, it would imply infallibilism about knowledge ascriptions. Still, the objection does point to a rather important point. There is an implicit folk theory of the meaning of 'knows', one according to which it does not denote justified true belief. I claim this folk theory is mistaken. It is odd to say that we can all be mistaken about the meanings of our words; it is odd to say that we can't make errors in word usage. I think the latter is the greater oddity, largely because I have a theory which explains how we can all make mistakes about meanings in our own language.

How can we make such mistakes? The short answer is that meanings ain't in the head. The long answer turns on the kind of tests on analyses I discussed in section two. The meaning of a predicate is a property in the sense described by Lewis (1983b)[5]: a set, or class, or plurality of possibilia. (That is, in general the meaning of a predicate is its intension.[6]) The interesting question is determining which property it is. In assigning a property to a predicate, there are two criteria we would like to follow. The first is that it validates as many as possible of our pre-theoretic beliefs. The second is that it is, in some sense, simple and theoretically important. How to make sense of this notion of simplicity is a rather complex matter. Lewis canvasses the idea that there is a primitive 'naturalness' of properties which measures simplicity and theoretical significance[7], and I will adopt this idea. Space restrictions prevent me going into greater detail concerning 'naturalness', but if something more definite is wanted, for the record I mean by it here just what Lewis means by it in the works previously cited.[8]

So, recapitulating what I said in section two, for any predicate t and property F, we want F meet two requirements before we say it is the meaning of t. We want this meaning assignment to validate many of our pre-theoretic intuitions (this is what we test for in tests one and two) and we want F to be reasonably natural (this is what we test for in tests three and four). In hard cases, these requirements pull in opposite directions; *the* meaning of t is the property which on balance does best. Saying 'knows' means 'justifiably truly believes' does not do particularly well on the first requirement. Gettier isolated a large class of cases where it goes wrong. But it does very well on the second, as it analyses knowledge in terms of a short list of simple and significant features. I claim that all its rivals don't do considerably better on the first, and arguably do much worse on the second. (There are considerations pulling either way here, as I note in section seven, but it is *prima facie* plausible that it does very well on the second, which

[5]The theory of meaning outlined here is deeply indebted to Lewis (1983b, 1984b, 1992).

[6]There are tricky questions concerning cointensional predicates, but these have fairly familiar solutions, which I accept. For ease of expression here I will ignore the distinction between properties and relations – presumably 'knows' denotes a relation, that is a set of ordered pairs.

[7]'Measures' may be inappropriate here. Plausibly a property is simple because it is natural.

[8]For more recent applications of naturalness in Lewis's work, see Langton and Lewis (1998, 2001)and Lewis (2001a).

is all that we consider for now.) That the JTB theory is the best trade-off is still a live possibility, even considering Gettier cases.

This little argument will be perfectly useless this theory of meaning (owing in all its essential features to Lewis) is roughly right. There are several reasons for believing it. First, it can account for the possibility of mistaken intuitions, while still denying the possibility that intuitions about meaning can be systematically and radically mistaken. This alone is a nice consequence, and not one which is shared by every theory of meaning on the market. Secondly, as was shown in sections one and two, it seems to make the right kinds of predictions about when meaning will diverge from intuitions about meaning.

Thirdly, it can account for the fact that some, but not all, disagreements about the acceptability of assertions are disputes about matters of fact, not matters of meaning. This example is from Cummins: "If a child, asked to use 'fair' in a sentence, says, "It isn't fair for girls to get as much as boys," we should suspect the child's politics, not his language" 1998, 120. This seems right; but if the child had said "It is fair that dreams are purple", we would suspect his language. Perhaps by 'fair' he means 'nonsensical' or something similar. A theory of meaning needs to account for this divergence, and for the fact that it is a vague matter when we say the problem is with the child's language, and when with his politics. In short, saying which disputes are disputes about facts (or values or whatever), and which about meanings, is a compulsory question for a theory of meaning.

The balance theory of meaning I am promoting can do this, as the following demonstration shows. This theory of meaning is determinedly individualistic. Every person has an idiolect determined by her dispositions to apply terms; a shared language is a collection of closely-enough overlapping idiolects. So the child's idiolect might differ from ours, especially if he uses 'fair' to mean 'nonsensical'. But if the idiolect differs in just how a few sentences are used, it is likely that the meaning postulate which does best at capturing his dispositions to use according to our *two* criteria, is the same as the meaning postulate which does best at capturing our dispositions to use. The reason is that highly natural properties are pretty thin on the ground; one's dispositions to use a term have to change quite a lot before they get into the orbit of a distinct natural property. So despite the fact that I allow for nothing more than overlapping idiolects, in practice the overlap is much closer to being exact than on most 'overlapping idiolect' theories.

With this, I can now distinguish which disputes are disputes about facts, and which are disputes about meaning. Given that there is a dispute, the parties must have different dispositions to use some important term. In some disputes, the same meaning postulate does best on balance at capturing the dispositions of each party. I say that here the parties mean the same thing by their words, and the dispute is a dispute about facts. In others, the difference will be so great that different meaning postulates do best at capturing the dispositions of the competing parties. In these cases, I say the dispute is a dispute about meaning.

Now, I can explain the intuition that the JTB theorist means something different to the rest of us by 'knows. That is, I can explain this intuition away. It seems a fair assumption that the reasonably natural properties will be evenly distributed throughout the space of possible linguistic dispositions. If this is right, then any change of usage beyond a certain magnitude will, on my theory, count as a change of meaning. And it is plausible to suppose the change I am urging to our usage, affirming rather than denying sentences like, "Smith knows Jones owns a Ford" is beyond that certain magnitude. But the assumption of even distribution of

the reasonably natural properties is false. That, I claim, is what the failure of the 'analysis of knowledge' merry-go-round to stop shows us. There are just no reasonably natural properties in the neighbourhood of our disposition to use 'knows'. If this is right, then even some quite significant changes to usage will not be changes in meaning, because they will not change which is the closest reasonably natural property to our usage pattern. The assumption that the reasonably natural properties are reasonably evenly distributed is plausible, but false. Hence the hunch that I am trying to change the meaning of 'knows' is plausible, but false.

The hypothesis that when we alter intuitions because of a theory we always change meanings, on the other hand, is not even plausible. When the ancients said "Whales are fish", or "The sun is not a star", they simply said false sentences. That is, they said that whales are fish, and believed that the sun is not a star. This seems platitudinous, but the 'use-change implies meaning-change' hypothesis would deny it.

It has sometimes been suggested to me that conceptual intuitions should be given greater privilege than other intuitions; that I am wrong to generalise from the massive fallibility of logical, ethical or semantic intuitions to the massive fallibility of conceptual intuitions. Since I am on much firmer ground when talking about these non-conceptual cases, if such an attack were justified it would severely weaken my argument. Given what has been said so far we should be able to see what is wrong with this suggestion. Consider a group of people who systematically assent to "If *A* then *B* implies if *B* then *A*." On this view these people are expressing a mistaken logical intuition, but a correct conceptual intuition. So their concept of 'implication' doesn't pick out implication, or at the very least doesn't pick out our concept of 'implication'. Now if we are in that group, this summary becomes incoherent, so this position immediately implies that we can't be mistaken about our logical intuitions. Further, we are no longer able to say that when these people say "If *A* then *B* implies if *B* then *A*," they are saying something false, because given the reference of 'implies' in their idiolect, this sentence expresses a true proposition. This is odd, but odder is to come. Assuming again we are in this group, it turns out to be vitally important in debates concerning philosophical logic to decide whether we are engaging in logical analysis or conceptual analysis. It might turn out a correct piece of conceptual analysis of 'implication' picks out a different relation to the correct implication relation we derive from purely logical considerations. If logical intuitions are less reliable than conceptual intuitions, as proposed, and assent to sentences like "If *A* then *B* implies if *B* then *A*" reveals simultaneously a logical and a conceptual intuition, this untenable conclusion seems forced. I conclude that conceptual intuitions are continuous with other intuitions, and should be treated in a similar way.

4 Keeping Conceptual Analysis

The following would be a bad way to respond to the worry that the JTB theory amounts to a change in the meaning of the word 'knows'. For the worry to have any bite, facts about the meaning of 'knows' will have to be explicable in terms of facts about the use of 'knows. But facts about use can only tell us about the beliefs of this community about knowledge, not what knowledge really is. Since different communities adopt different standards for knowledge, we should only trust ours over theirs if (a) we have special evidence that our is correct or (b) we are so xenophobic that we trust ours simply because it is ours. "Many of us care very much whether or cognitive processes lead to beliefs that are true, or give us power over nature, or lead to

happiness. But only those with a deep and free-floating conservatism in matters epistemic will care whether their cognitive processes are sanctioned by the evaluative standards that happen to be woven into our language" (Stich, 1988, 109). "The intuitions and tacit knowledge of the man or woman in the street are quite irrelevant. The theory seeks to say what [knowledge] really is, not what folk [epistemology] takes it to be" (Stich, 1992, 252)[9]. Facts about use can only give us the latter, so they are not what are relevant to my inquiry.

Stich takes this to be a general reason for abandoning conceptual analysis. Now while I think, and have argued above, that conceptual analysis need not slavishly follow intuition, I do not think that we should abandon it altogether. Stich's worry seems to be conceptual analysis can only tell us about our words, not about our world. But is this kind of worry coherent? Can we say what will be found when we get to this real knowledge about the world? Will we be saying, "This belief of Smith's shouldn't be called knowledge, but really it is"? We need to attend to facts about the meaning of 'knows' in order to define the target of our search. If not, we have no way to avoid incoherencies like this one.

To put the same point another way, when someone claims to find this deep truth about knowledge, why should anyone else care? She will say, "Smith really knows that Jones owns a Ford, but I don't mean what everyone else means by 'knows'." Why is this any more interesting than saying, "Smith really is a grapefruit, but I don't mean what everyone else means by 'grapefruit'"? If she doesn't use words in the way that we do, we can ignore what she says about our common word usage. Or at least we can ignore it until she (or one of her colleagues) provides us with a translation manual. But to produce a translation manual, or to use words the way we do, she needs to attend to facts about our meanings. Again, incoherence threatens if she doesn't attend to these facts but claims nevertheless to be participating in a debate with us. These points are all to be found in Chapter 2 of Jackson (1998).

An underlying assumption of the first reply is that there is a hard division between facts about meaning and facts about the world at large; that a principle like: *No 'is' from a 'means'* holds. This principle is, however, mistaken. All instances of the following argument pattern, where t ranges over tokenings of referring terms, are valid.

P1: t refers unequivocally to α.
P2: t refers unequivocally to β.
C: $\alpha = \beta$

For example, from the premise that 'POTUS' refers unequivocally to the President of the United States, and the premise that 'POTUS' refers unequivocally to Bush, we can validly infer that Bush is President of the United States. Since P1 and P2 are facts about meaning, and C is a fact about the world, any principle like *No 'is' from a 'means'* must be mistaken. So this worry about how much we can learn from conceptual analysis, from considerations of meaning, is mistaken.

I call this inference pattern the R-inference. That the R-inference is valid doesn't just show Stich's critique rests on the false assumption *No 'is' from a 'means'*. It can be used to provide

[9]The paper from which this quote is drawn is about the content of mental states, so originally it had 'mental representation' for 'knowledge' and 'psychology' for 'epistemology'. But I take it that (a) this isn't an unfair representation of Stich's views and (b) even if it is, it is an admirably clear statement of the way many people feel about the use of intuitions about possible cases, and worth considering for that reason alone.

a direct response to his critique. The problem is meant to be that conceptual analysis, the method of counterexamples, can at best provide us with claims like: 'knows' refers to the relation *justifiably truly believes*. We want to know facts about knowledge, not about the term 'knows', so the conceptual analyst seems to have been looking in the wrong place. But it is a platitude that 'knows' refers to the relation *knows*. I call such platitudes, that '*t*' refers to *t*, instances of the R-schema[10]. We can use the R-schema together with the R-inference to get the kind of conclusion our opponents are looking for.

P1: 'Knowledge' refers unequivocally to the relation *justifiably truly believes*.
P2: 'Knowledge' refers unequivocally to the relation *knows*.
C: The relation *knows* is the relation *justifiably truly believes*.

More colloquially, the conclusion says that knowledge is justified true belief. Everyone agrees (I take it) that conceptual analysis could, in principle, give us knowledge of facts of the form of P1. So the opponents of conceptual analysis must either deny P2, or deny that C follows from P1 and P2. In other words, for any such argument they must deny that the R-schema is true, or that the R-inference is valid. I hope the reader will agree that neither option looks promising.

5 Against the Psychologists

Someone excessively impressed by various results in the psychological study of concepts may make the following objection to the theory of meaning here proffered. "Why think that we should prefer short lists of necessary and sufficient conditions? This seems like another one of those cases where philosophers take their aesthetic preferences to be truth-indicative, much like the 'taste for desert landscapes' argument. Besides, haven't psychologists like Eleanor Rosch shown that our concepts don't have simple necessary and sufficient conditions? If that's right, your argument falls down in several different places."

Strictly speaking, my preference is not just for short lists of necessary and sufficient conditions. But it is, for reasons set out more fully in the next section, for short theories that fit the meaning of some term into a network of other properties. And my argument would fall down if there was no reason to prefer such short theories. And, of course, short lists of necessary and sufficient conditions are paradigmatically short theories. One reason I prefer the JTB analysis to its modern rivals is its brevity. Some of the reasons for preferring short lists are brought out by considering the objections to this approach developed by psychologists. I'll just focus on one of the experiments performed by Rosch and Mervis, the points I make can be generalised.

Rosch and Mervis (1975) claim that "subjects rate superordinate semantic categories as having few, if any, attributes common to all members." (p. 20) (A superordinate semantic category is one, like 'fruit', which has other categories, like 'apple', 'pear' and 'banana', as sub-categories.) Here's the experiment they ran to show this. For each of six superordinate categories ('furniture', 'fruit', 'weapon', 'vegetable', 'vehicle' and 'clothing') they selected twenty category members. So for 'fruit' the members ranged from 'orange' and 'apple' to 'tomato'

[10] (Horwich, 1999, 115-130) discusses similar schema, noting that instances involving words in foreign languages, or indexical expressions, will not be platitudinous. He also notes a way to remove the presumption that there is such a thing as knowledge, by stating the schema as $\forall x$ ('knowledge' refers to x iff knowledge = x). For ease of expression I will stick with the simpler formulation in the text.

and 'olive'. They then asked a range of subjects to list the attributes they associated with some of these 120 category members. Each subject was presented with six members, one from each category, and for each member had a minute and a half to write down its salient attributes.

> [F]ew attributes were given that were true of all twenty members of the category – for four of the categories there was only one such item; for two of the categories, none. Furthermore, the single attribute that did apply to all members, in three cases was true of many items besides those within that superordinate (for example, "you eat it" for fruit). Rosch and Mervis (1975)

They go on to conclude that the superordinate is not defined by necessary and sufficient conditions, but by a 'family resemblance' between members. This particular experiment was taken to confirm that the number of attributes a member has with other members of the category is correlated with a previously defined measure of prototypicality.[11] They claim that the intuition, commonly held amongst philosophers, that there must be some attribute in common to all the members, is explicable by the fact that the highly prototypical members of the category all do share quite a few attributes in common, ranging from 3 attributes in common to the highly prototypical vegetables, to 36 for the highly prototypical vehicles.

One occasionally hears people deride the assumption that there are necessary and sufficient conditions for the application of a term, as if this was the most preposterous piece of philosophy possible. Really, this assumption is no more than the assumption that dictionaries can be written, and without any reason to think otherwise, seems perfectly harmless. Perhaps, though, the Rosch and Mervis experiments provide a reason to think otherwise, a reason for thinking that the conditions of applicability for terms like 'fruit', 'weapon', and perhaps 'knowledge' are Wittgensteinian family resemblance conditions, rather than short lists of necessary and sufficient conditions, the kinds of conditions that fill traditional dictionaries.

When we look closely, we see that the experiments do not show this at all. One could try and knock any such argument away by claiming the proposal is incoherent. The psychologists claim that there are no necessary and sufficient conditions for being a weapon, but something is a weapon iff it bears a suitable resemblance to paradigmatic weapons. In one sense, bearing a suitable resemblance to a paradigmatic weapon is a condition, so it looks like we just have a very short list of necessary and sufficient conditions, a list of length one. (Jackson, 1998, 61) makes a similar point in response to Stich's invocation of Rosch's experiments. This feels like it's cheating, so I'll move onto other objections. I'll explain below just why it feels like cheating.

Philosophers aren't particularly interested in terms like 'weapon', so these experiments only have *philosophical* interest if the results can be shown to generalise to terms philosophers care about. In other words, if can be shown that terms like 'property', 'justice', 'cause' and particularly 'knows' are cluster concepts, or family resemblance terms. But there is a good reason to think this is false. As William Ramsey (1998) notes, if F refers to a cluster concept, then for any proposed list of necessary and sufficient properties for F-hood, it should be easy to find an individual which is an F but which lacks some of these properties. To generate such an example, just find an individual which lacks one of the proposed properties, but which has several other properties from the cluster. It should be harder to find an individual which has the

[11] In previous work they had done some nice experiments aimed at getting a grip on our intuition that apples are more prototypical exemplars of fruit than olives are.

properties without being an *F*. If the proposed analysis is even close to being right, then having these conditions will entail having enough of the cluster of properties that are constitutive of *F*-hood to be an *F*. Note, for example, that all of the counterexamples Wittgenstein (1953) lists to purported analyses of 'game' are cases where something is, intuitively, a game but which does not satisfy the analysis. If game is really a cluster concept, this is how things should be. But it is not how things are with knowledge; virtually all counterexamples, from Gettier on, are cases which are intuitively not cases of knowledge, but which satisfy the proposed analysis. This is good evidence that even if some terms in English refer to cluster concepts, 'knows' is not one of them.

Secondly, Rosch and Mervis's conclusions about the nature of the superordinate categories makes some rather mundane facts quite inexplicable. In this experiment the subjects weren't told which category each member was in, but for other categories they were. Imagine, as seems plausible, one of the subjects objected to putting the member in that category. Many people, even undergraduates, don't regard olives and tomatoes as fruit. ("Fruit on pasta? How absurd!") When the student asks why is this thing called a fruit, other speakers can provide a response. It is not a brute fact of language that tomatoes are fruit. It is not just by magic that we happened to come to a shared meaning for fruit that includes tomatoes, and that if faced with a new kind of object, we would generally agree about whether it is a fruit. It is because we know how to answer such questions. This answer to the *Why is it called 'fruit'?* question had better be a sufficient condition for fruitness. If not, the subject is entitled to ask why having that property makes it a fruit. And unless there are very many possible distinct answers to this question, which seems very improbable, there will be a short list of necessary and sufficient conditions for being a fruit. But for this example, at least, 'fruit' was relatively arbitrary, so there will be a short list of necessary and sufficient conditions for being an *F*, for pretty much any *F*.

Thirdly, returning to 'fruit', we can see that Rosch and Mervis's experiments could not possibly show that many superordinate predicates in English are cluster concepts. For they would, if successful, show that 'fruit' is a cluster concept, and it quite plainly is not. So by *modus tollens*, there is something wrong with their methodology. Some of the other categories they investigate, particularly 'weapon' and 'furniture' *might* be relatively cluster-ish, in a sense to be explained soon, but not 'fruit'. As the OED says, a fruit is "the edible product of a tree, shrub or other plant, consisting of the seed and its envelope." If nothing like this is right, then we couldn't explain to the sceptical why we call tomatoes, olives and so on fruit.

So the conclusion that philosophically significant terms are likely to be cluster concepts is mistaken. To close, I note one way the cluster concept view could at least be coherent. Many predicates do have necessary and sufficient conditions for their applicability, just as traditional conceptual analysis assumed. In other words, they have analyses. However, any analysis must be in words, and sometimes the words needed will refer to quite recherche properties. The properties in the analysans may, that is, be significantly less natural than the analysandum.

In some contexts, we only consider properties that are above a certain level of naturalness. If I claim two things say my carpet and the Battle of Agincourt, have nothing in common, I will not feel threatened by an objector who points out that they share some gruesome, gerrymandered property, like being elements of {my carpet, the Battle of Agincourt}. Say that the best analysis of *F*-hood requires us to use predicates denoting properties which are below the contextually defined border between the 'natural enough' and 'too gruesome to use'. Then there will be a sense in which there is no analysis of *F* into necessary and sufficient conditions;

just the sense in which my carpet and the Battle of Avignon have nothing in common. Jackson's argument feels like a cheat because he just shows that there will be necessary and sufficient conditions for any concept provided we are allowed to use gruesome properties, but he makes it sound like this proviso is unnecessary. If Rosch and Mervis's experiments show anything at all, it is that this is true of some common terms in some everyday-ish contexts. In particular, if we restrict our attention to the predicates that might occur to us within ninety seconds (which plausibly correlates well with some level of naturalness), very few terms have analyses. Thus far, Rosch and Mervis are correct. They go wrong by projecting truths of a particular context to all contexts.

6 In defence of analysis

In the previous section I argued that various empirical arguments gave us no reason to doubt that 'knows' will have a short analysis. In this section we look at various philosophical arguments to this conclusion. One might easily imagine the following objection to what has been claimed so far. At best, the above reasoning shows that if 'knows' has a short analysis, then the JTB analysis is correct, notwithstanding the intuitions provoked by Gettier cases. But there is little reason to think English terms have analyses, as evidenced by the failure of philosophers to analyse even one interesting term, and particular reasons to think that 'knows' does not have an analysis. These reasons are set out by (Williamson, 2000, Ch. 3), who argues, by appeal to intuitions about a particular kind of case, that there can be no analysis of 'knows' into independent clauses, one of which describes an internal state of the agent and the other of which describes an external state of the agent. This does not *necessarily* refute the JTB analysis, since the concepts of justification and belief in use may be neither internal nor external in Williamson's sense. And if we are going to revise intuitions about the Gettier cases, we may wish to revise intuitions about Williamson's cases as well, though here it is probably safest to *not* do this, because it is unclear just what philosophical benefit is derived from this revision. In response to these arguments I will make two moves: one defensive and one offensive. The defensive move is to distinguish the assumptions made here about the structure of the meaning of 'knows', and show how these assumptions do not have some of the dreadful consequences suggested by various authors. The offensive move, with which we begin, is to point out the rather unattractive consequences of *not* making these assumptions about the structure of the meaning of 'knows'.

In terms of the concept of naturalness used above, the relation denoted by 'knows' might fall into one of three broad camps:

(a) It might be rather unnatural;
(b) It might be fairly natural in virtue of its relation to other, more natural, properties; or
(c) It might be a primitive natural property, one that does not derive its naturalness from anything else.

My preferred position is (b). I think that the word 'knows', like every other denoting term in English, denotes something fairly natural. And I don't think there are any primitively natural properties or relations in the vicinity of the denotation of this word, so it must derive its naturalness from its relation to other properties or relations. If this is so, we can recover some of the structure of its meaning by elucidating those relationships. If it is correct, that is exactly what

I think the JTB theory does. This is not to say that justification, truth or belief are themselves primitively natural properties, but rather that we can make some progress towards recovering the source of the naturalness of knowledge via its decomposition into justification, truth and belief. But before investigating the costs of (b), let us look at the costs of (a) and (c).

I think we can dispense with (c) rather quickly. It would be surprising, to say the least, if knowledge was a primitive relation. That X knows that p can hardly be one of the foundational facts that make up the universe. If X knows that p, this fact obtains in virtue of the obtaining of other facts. We may not be able to tell exactly what these facts are in general, but we have fairly strong opinions about whether they obtain or not in a particular case. This is why we are prepared to say whether or not a character knows something in a story, perhaps a philosophical story, without being told exactly that. We see the facts in virtue of which the character does, or does not, know this. This does not *conclusively* show that knowledge is not a primitively natural property. Electrical charge presumably is a primitively natural property, yet sometimes we can figure out the charge of an object by the behaviour of other objects. For example, if we know it is repulsed by several different negatively charged things, it is probably negatively charged. But in these cases it is clear our inference is from some facts to other facts that are inductively implied, not to facts that are constituted by the facts we know. (Only a rather unreformed positivist would say that charge is *constituted* by repulsive behaviour.) And it does not at all feel that in philosophical examples we are inductively (or abductively) inferring whether the character knows that p.

The more interesting question is whether (a) might be correct. This is, perhaps surprisingly, consistent with the theory of meaning advanced above. I held, following Lewis, that the meaning of a denoting term is the most natural object, property or relation that satisfies most of our usage dispositions. It is possible that the winner of this contest will itself be quite unnatural. This is what happens all the time with vague terms, and indeed it is what causes, or perhaps constitutes, their vagueness. None of the properties (or relations) that we may pick out by 'blue' is much more natural than several other properties (or relations) that would do roughly as well at capturing our usage dispositions, were they the denotation of 'blue'.[12] And indeed none of these properties (or relations) are particularly natural; they are all rather arbitrary divisions of the spectrum. The situation is possibly worse when we consider what Theodore Sider (2001c) calls maximal properties. A property F is maximal iff things that massively overlap an F are not themselves an F. So *being a coin* is maximal, since large parts of a coin, or large parts of a coin fused with some nearby atoms outside the coin, are not themselves coins. Sider adopts the following useful notation: something is an F^* iff it is suitable to be an F in every respect save that it may massively overlap an F. So a coin* is a piece of metal (or suitable substance) that is (roughly) coin-shaped and is (more or less) the deliberate outcome of a process designed to produce legal tender. Assuming that any collection of atoms has a fusion, in the vicinity of any coin there will be literally trillions of coin*s. At most one of these will be a coin, since coins do not, in general, overlap. That is, the property *being a coin* must pick out exactly one of these coin*s. Since the selection will be ultimately arbitrary, this property is not very natural. There are just no natural properties in the area, so the denotation of 'coin' is just not natural.

[12] I include the parenthetical comments here so as not to prejudge the question of whether colours are properties or relations. It seems unlikely to me that colours are relations, either the viewers or environments, but it is not worth quibbling over this here.

These kind of considerations show that option (a) is a live possibility. But they do not show that it actually obtains. And there are several contrasts between 'knows', on the one hand, and 'blue' and 'coin' on the other, which suggest that it does not obtain. First, we do not take our word 'knows' to be as indeterminate as 'blue' or 'coin', despite the existence of some rather strong grounds for indeterminacy in it. Secondly, we take apparent disputes between different users of the word 'knows' to be genuine disputes, ones in which at most one side is correct, which we do not necessarily do with 'blue' and 'coin'. Finally, we are prepared to use the relation denoted by 'knows' in inductive arguments in ways that seem a little suspect with genuinely unnatural relations, as arguably evidenced by our attitudes towards 'coin' and 'blue'. Let's look at these in more detail.

If we insisted that the meaning of 'knows' must validate *all* of our dispositions to use the term, we would find that the word has no meaning. If we just look at intuitions, we will find that our intuitions about 'knows' are inconsistent with some simple known facts. (Beliefs, being regimented by reflection, *might* not be inconsistent, depending on how systematic the regimentation has been.) For example, the following all seem true to many people.

(1) Knowledge supervenes on evidence: if two people (not necessarily in the same possible world) have the same evidence, they know the same things.
(2) We know many things about the external world.
(3) We have the same evidence as some people who are the victims of massive deception, and who have few true beliefs about their external world.
(4) Whatever is known is true.

These are inconsistent, so they cannot all be true. We could take any three of these as an argument for the negation of the fourth, though probably the argument from (1) (2) and (3) to the negation of (4) is less persuasive than the other three such arguments. I don't want to adjudicate here which such argument is sound. All I want to claim here is that there is a fact of the matter about which of these arguments is sound, and hence about which of these four claims is false. If two people are disagreeing about which of these is false, at most one of them is right, and the other is wrong. If 'knows' denoted a rather unnatural relation, there would be little reason to believe these things to be true. Perhaps by more carefully consulting intuitions we could determine that one of them is false by seeing that it had the weakest intuitive pull. If we couldn't do this, it would follow that in general there was no fact of the matter about which is false, and if someone wanted to use 'know' in their idiolect so that one particular one of these is false, there would be no way we could argue that they were wrong. It is quite implausible that this is what should happen in such a situation. It is more plausible that the dispute should be decided by figuring out which group of three can be satisfied by a fairly natural relation. This, recall, is just how we resolve disputes in many other areas of philosophy, from logic to ethics. If there is no natural relation eligible to be the meaning of 'knows', then probably this dispute has no resolution, just like the dispute about what 'mass' means in Newtonian mechanics.[13]

The above case generalises quite widely. If one speaker says that a Gettier case is a case of knowledge and another denies this (as Stich assures us actually happens if we cast our linguistic

[13]Note that in that dispute the rivals are quite natural properties, but seem to be matched in their naturalness. In the dispute envisaged here, the rivals are quite unnatural, but still seem to be matched. For more on 'mass', see Field (1973).

net wide enough) we normally assume that one of them is making a mistake. But if 'knows' denotes something quite unnatural, then probably each is saying something true in her own idiolect. Each party may make other mistaken claims, that for example what they say is also true in the language of all their compatriots, but in just making these claims about knowledge they would not be making a mistake. Perhaps there really is no fact of the matter here about who is right, but thinking so would be a major change to our common way of viewing matters, and hence would be a rather costly consequence of accepting option (a). Note here the contrast with 'blue' and 'coin'. If one person adopts an idiosyncratic usage of 'blue' and 'coin', one on which there are determinate facts about matters where, we say, there are none, the most natural thing to say is that they are using the terms differently to us. If they insist that it is part of their intention in using the terms to speak the same way as their fellows we may (but only may) revise this judgement. But in general there is much more inclination to say that a dispute over whether, say, a patch is blue is merely verbal than to say this about a dispute over whether X knows that p.

Finally, if knowledge was a completely unnatural relation, we would no more expect it to play a role in inductive or analogical arguments than does grue, but it seems it can play such a role. One might worry here that blueness also plays a role in inductive arguments, as in: The sky has been blue the last n days, so probably it will be blue tomorrow. If blueness is not natural, this might show that unnatural properties can play a role in inductive arguments. But what is really happening here is that there is, implicitly, an inductive argument based on a much narrow colour spectrum, and hence a much more natural property. To see this, note that we would be just as surprised tomorrow if the sky was navy blue, or perhaps of the dominant blue in Picasso's blue period paintings, as if it were not blue at all.

So there are substantial costs to (a) and (c). Are there similar costs to (b)? If we take (b) to mean that there is a decomposition of the meaning of 'knows' into conditions, expressible in English, which we can tell *a priori* are individually necessary and jointly sufficient for knowledge, and such that it is also *a priori* that they represent natural properties, then (b) would be wildly implausible. To take just one part of this, Williamson (2000) notes it is clear that there are some languages in which such conditions cannot be expressed, so perhaps English is such a language too. And if this argument for 'knows' works it presumably works for other terms, like 'pain', but it is hard to find such an *a priori* decomposition of 'pain' into more natural properties. Really, all (b) requires is that there be some connection, perhaps only discoverable *a posteriori*, perhaps not even humanly comprehensible, between knowledge and other more primitively natural properties. These properties need not be denoted by any terms of English, or any other known language.

Most importantly, this connection need not be a decomposition. If knowledge is the most general factive mental state, as Williamson proposes, and being factive and being a mental state are natural properties, then condition (b) will be thereby satisfied. If knowledge is the norm of assertion, as Williamson also proposes, then that could do as the means by which knowledge is linked into the network of natural properties. This last assumes that *being an assertion* is a natural property, and more dangerously that norms as natural, but these are relatively plausible assumptions in general. In neither case do we have a factorisation, in any sense, of knowledge into constituent properties, but we do have, as (b) requires, a means by which knowledge is linked into the network of natural properties. It is quite plausible that for every term which,

unlike 'blue' and 'coin' are not excessively vague and do not denote maximal properties, something like (b) is correct. Given the clarifications made here to (b), this is consistent with most positions normally taken to be anti-reductionist about those terms, or their denotata.

7 Naturalness and the JTB theory

I have argued here that the following argument against the JTB theory is unsound.

P1. The JTB theory says that Gettier cases are cases of knowledge.
P2. Intuition says that Gettier cases are not cases of knowledge.
P3. Intuition is trustworthy in these cases.
C. The JTB theory is false.

The objection has been that P3 is false in those cases where following intuition slavishly would mean concluding that some common term denoted a rather unnatural property while accepting deviations from intuition would allow us to hold that it denoted a rather natural property. Peter Klein (in conversation) has suggested that there is a more sophisticated argument against the JTB theory that we can draw out of the Gettier cases. Since this argument is a good illustration of the way counterexamples should be used in philosophy, I'll close with it.

Klein's idea, in effect, is that we can use Gettier cases to argue that *being a justified true belief* is not a natural property, and hence that P3 is after all true. Remember that P3 only fails when following intuition too closely would lead too far away from naturalness. If *being a justified true belief* is not a natural property to start with, there is no great danger of this happening. What the Gettier cases show us, goes the argument, is that there are two ways to be a justified true belief. The first way is where the belief is justified in some sense because it is true. The second way is where it is quite coincidental that the belief is both justified and true. These two ways of being a justified true belief may be natural enough, but the property *being a justified true belief* is just the disjunction of these two not especially related properties.

I think this is, at least, a *prima facie* compelling argument. There are, at least, three important points to note about it. First, this kind of reasoning does not obviously generalise. Few of the examples described in Shope (1983) could be used to show that some target theory in fact made knowledge into a disjunctive kind. The second point is that accepting this argument is perfectly consistent with accepting everything I said above against the (widespread) uncritical use of appeal to intuition. Indeed, if what I said above is broadly correct then this is just the kind of reasoning we should be attempting to use when looking at fascinating counterexamples. Thirdly, if the argument works it shows something much more interesting than just that the JTB theory is false. It shows that naturalness is not always transferred to a conjunctive property by its conjuncts.

I assume here that *being a justified belief* and *being a true belief* are themselves natural properties, and *being a justified true belief* is the conjunction of these. The only point here that seems possibly contentious is that *being a true belief* is not natural. On some forms of minimalism about truth this may be false, but those forms seem quite implausibly strong. Remember that saying *being a true belief* is natural does not imply that has an analysis – truth might be a primitively natural component of this property. And remember also that naturalness is intensional rather than hyperintensional. If all true beliefs correspond with reality in a

suitable way, and *corresponding with reality in that way* is a natural property, then so is *being a true belief*, even if truth of belief cannot be explained in terms of correspondence.

This is a surprising result, because the way naturalness was originally set up by Lewis suggested that it would be transferred to a conjunctive property by its conjuncts. Lewis gave three accounts of naturalness. The first is that properties are perfectly natural in virtue of being co-intensive with a genuine universal. The third is that properties are natural in virtue of the mutual resemblance of their members, where resemblance is taken to be a primitive. On either account, it seems that whenever *being F* is natural, and so is *being G*, then *being F and G* will be natural.[14] The second account, if it can be called that, is that naturalness is just primitive. If the Gettier cases really do show that *being a justified true belief* is not natural, then they will have shown that we have to fall back on just this account of naturalness.

[14] I follow Armstrong (1978) here in assuming that there are conjunctive universals.

Morality, Fiction and Possibility

1 Four Puzzles

Several things go wrong in the following story.

> *Death on a Freeway*
> Jack and Jill were arguing again. This was not in itself unusual, but this time they were standing in the fast lane of I-95 having their argument. This was causing traffic to bank up a bit. It wasn't significantly worse than normally happened around Providence, not that you could have told that from the reactions of passing motorists. They were convinced that Jack and Jill, and not the volume of traffic, were the primary causes of the slowdown. They all forgot how bad traffic normally is along there. When Craig saw that the cause of the bankup had been Jack and Jill, he took his gun out of the glovebox and shot them. People then started driving over their bodies, and while the new speed hump caused some people to slow down a bit, mostly traffic returned to its normal speed. So Craig did the right thing, because Jack and Jill should have taken their argument somewhere else where they wouldn't get in anyone's way.

The last sentence raises a few related puzzles. Intuitively, it is not true, even in the story, that Craig's murder was morally justified. What the narrator tells us here is just false. That should be a little surprising. We're being told a story, after all, so the storyteller should be an authority on what's true in it. Here we hearers get to rule on which moral claims are true and false, not the author. But usually the author gets to say what's what. The action takes place in Providence, on Highway 95, just because the author says so. And we don't reject those claims in the story just because no such murder has ever taken place on Highway 95. False claims can generally be true in stories. Normally, the author's say so is enough to make it so, at least in the story, even if what is said is really false. The first puzzle, the **alethic** puzzle, is why authorial authority breaks down in cases like *Death on the Freeway*. Why can't the author just make sentences like the last sentence in *Death* true in the story by saying they are true? At this stage I won't try and give a more precise characterisation of which features of *Death* lead to the break down of authorial authority, for that will be at issue below.

The second puzzle concerns the relation between fiction and imagination. Following Kendall Walton (1990), it is common to construe fictional works as invitations to imagine. The author requests, or suggests, that we imagine a certain world. In *Death* we can follow along with the author for most of the story. We can imagine an argument taking place in peak hour on Highway 95. We can imagine this frustrating the other drivers. And we can imagine one of those

† Penultimate draft only. Please cite published version if possible. Final version published in *Philosophers' Imprint* vol. 4, number 3. I've spoken to practically everyone I know about the issues here, and a full list of thanks for useful advice, suggestions, recommendations, criticisms, counterexamples and encouragement would double the size of the paper. If I thank philosophy departments rather than all the individuals in them it might cut the size a little, so thanks to the departments at Brown, UC Davis, Melbourne, MIT and Monash. Thanks also to Kendall Walton, Tamar Gendler and two referees for *Philosophers' Imprint*. The most useful assistance came from Wolfgang Schwarz and especially Tyler Doggett, without whose advice this could never have been written, and to George Wilson, who prevented me from (keeping on) making a serious error of over-generalisation.

drivers retaliating with a loaded gun. What we cannot, or at least do not, imagine is that this retaliation is morally justified. There is a limit to our imaginative ability here. We refuse, fairly systematically, to play along with the author here. Call this the **imaginative** puzzle. Why don't we play along in cases like *Death*? Again, I won't say *for now* which cases are like *Death*.

The third puzzle concerns the phenomenology of *Death* and stories like it. The final sentence is striking, jarring in a way that the earlier sentences are not. Presumably this is closely related to the earlier puzzles, though I'll argue below that the cases that generate this peculiar reaction are not identical with cases that generate alethic or imaginative puzzles. So call this the **phenomenological** puzzle.

Finally, there is a puzzle that David Hume (1757) first noticed. Hume suggested that artistic works that include morally deviant claims, moral claims that wouldn't be true were the descriptive aspects of the story true, are thereby aesthetically compromised. Why is this so? Call that the **aesthetic** puzzle. I will have nothing to say about that puzzle here, though hopefully what I have to say about the other puzzles will assist in solving it.

I'm going to call sentences that raise the first three puzzles **puzzling** sentences. Eventually I'll look at the small differences between those three puzzles, but for now we'll focus on what they have in common. The puzzles, especially the imaginative puzzle, have become quite a focus of debate in recent years. The aesthetic puzzle is raised by David Hume (1757), and is discussed by Kendall Walton (1994) and Richard Moran (1995). Walton and Moran also discuss the imaginative and alethic puzzles, and they are the focus of attention in recent work by Tamar Szabó Gendler (2000), Gregory Currie (2002) and Stephen Yablo (2002). My solution to the puzzles is best thought of as a development of some of Walton's 'sketchy story' (to use his description). Gendler suggests one way to develop Walton's views, and shows it leads to an unacceptable solution, because it leads to mistaken predictions. I will argue that there are more modest developments of Walton's views that don't lead to so many predictions, and in particular don't lead to *mistaken* predictions, but which still say enough to solve the puzzles.

2 The Range of the Puzzles

As Walton and Yablo note, the puzzle does not only arise in connection with thin moral concepts. But it has not been appreciated how widespread the puzzle is, and getting a sense of this helps us narrow the range of possible solutions.

Sentences in stories attributing **thick moral concepts** can be puzzling. If my prose retelling of *Macbeth* included the line "Then the cowardly Macduff called on the brave Macbeth to fight him face to face," the reader would not accept that in the story Macduff was a coward. If my retelling of *Hamlet* frequently described the young prince as decisive, the reader would struggle to go along with me imaginatively. Try imagining Hamlet doing exactly what he does, and saying exactly what he says, and thinking what he thinks, but always decisively. For an actual example, it's easy to find the first line in Bob Dylan's *Ballad of Frankie Lee and Judas Priest*, that the titular characters 'were the best of friends' puzzling in the context of how Frankie Lee treats Judas Priest later in the song. It isn't too surprising that the puzzle extends to the thick moral concepts, and Walton at least doesn't even regard these as a separate category.

More interestingly, any kind of **evaluative** sentence can be puzzling. Walton and Yablo both discuss sentences attributing aesthetic properties. (Yablo, 2002, 485) suggests that a story in which the author talks about the sublime beauty of a monster truck rally, while complaining

about the lack of aesthetic value in sunsets, is in most respects like our morally deviant story. The salient aesthetic claims will be puzzling. Note that we *are* able to imagine a community that prefers the sight of a 'blood bath death match of doom' (to use Yablo's evocative description) to sunsets over Sydney Harbour and it could certainly be true in a fiction that such attitudes were commonplace. But that does not imply that those people could be *right* in thinking the trucks are more beautiful. (Walton, 1994, 43-44) notes that sentences describing jokes that are actually unfunny as being funny will be puzzling. We get to decide what is funny, not the author.

Walton and Yablo's point here can be extended to **epistemic evaluations**. Again it isn't too hard to find puzzling examples when we look at attributions of rationality or irrationality.

> *Alien Robbery*
> Sam saw his friend Lee Remnick rushing out of a bank carrying in one hand a large bag with money falling out of the top and in the other hand a sawn-off shotgun. Lee Remnick recognised Sam across the street and waved with her gun hand, which frightened Sam a little. Sam was a little shocked to see Lee do this, because despite a few childish pranks involving stolen cars, she'd been fairly law abiding. So Sam decided that it wasn't Lee, but really a shape-shifting alien that looked like Lee, that robbed the bank. Although shape-shifting aliens didn't exist, and until that moment Sam had no evidence that they did, this was a rational belief. False, but rational.

The last two sentences of *Alien Robbery* are fairly clearly puzzling.

So far all of our examples have involved normative concepts, so one might think the solution to the puzzle will have something to do with the distinctive nature of normative concepts, or with their distinctive role in fiction. Indeed, Gendler's and Currie's solutions have just this feature. But sentences that seem somewhat removed from the realm of the normative can still be puzzling. (It is of course contentious just where the normative/non-normative barrier lies. Most of the following cases will be regarded as involving normative concepts by at least some philosophers. But I think few people will hold that *all* of the following cases involve normative concepts.)

Attributions of **mental states** can, in principle, be puzzling. If I retell *Romeo and Juliet*, and in this say 'Although he believed he loved Juliet, and acted as if he did, Romeo did not really love Juliet, and actually wanted to humiliate her by getting her to betray her family', that would I think be puzzling. This example is odd, because it is not obviously *impossible* that Romeo could fail to love Juliet even though he thought he loved her (people are mistaken about this kind of thing all the time) and acted as if he did (especially if he was trying to trick her). But given the full detail of the story, it is impossible to imagine that Romeo thought he had the attitudes towards Juliet he is traditionally thought to have, and he is mistaken about this.

Attributions of **content**, either mental content or linguistic content, can be just as puzzling. The second and third sentences in this story are impossible to imagine, and false even in the story.

Cats and Dogs
Rhodisland is much like a part of the actual world, but with a surprising difference. Although they use the word 'cat' in all the circumstances when we would (i.e. when they want to say something about cats), and the word 'dog' in all the circumstances we would, in their language 'cat' means dog and 'dog' means cat. None of the Rhodislanders are aware of this, so they frequently say false things when asked about cats and dogs. Indeed, no one has ever known that their words had this meaning, and they would probably investigate just how this came to be in some detail, if they knew it were true.

A similar story can be told to demonstrate how claims about mental content can be puzzling. Perhaps these cases still involve the normative. Loving might be thought to entail special obligations and Kripke (1982) has argued that content is normative. But we are clearly moving away from the moral, narrowly construed.

Stephen Yablo recently suggested that certain **shape** predicates generate imaginative resistance. These predicates are meant to be special categories of a broader category that we'll discuss further below. Here's Yablo's example.

Game Over
They flopped down beneath the giant maple. One more item to find, and yet the game seemed lost. Hang on, Sally said. It's staring us in the face. This is a *maple* tree we're under. She grabbed a five-fingered leaf. Here was the oval they needed! They ran off to claim their prize. (Yablo, 2002, 485, title added)

There's a potential complication in this story in that one might think that it's metaphysically impossible that *maple* trees have ovular leaves. That's not what is meant to be resisted, and I don't think is resisted. What is resisted is that maple leaves have their distinctive five-fingered look, that the shape of the leaf Sally collects is like *that* (imagine I demonstrate a maple leaf here) and that its shape be an *oval*.

Fewer people may care about the next class of cases, or have clear intuitions about them, but if one has firm **ontological** beliefs, then deviant ontological claims can be puzzling. I'm a universalist about mereology, at least with respect to ordinary concrete things, so I find many of the claims in this story puzzling.

Wiggins' World
The Hogwarts Express was a very special train. It had no parts at all. Although you'd be tempted to say that it had carriages, an engine, seats, wheels, windows and so on, it really was a mereological atom. And it certainly had no temporal parts - it wholly was wherever and whenever it was. Even more surprisingly, it did not enter into fusions, so when the Hogwarts Local was linked to it for the first few miles out of Kings Cross, there was no one object that carried all the students through north London.

I think that even in fictions any two concrete objects have a fusion. So the Hogwarts Express and the Hogwarts Local have a fusion, and when it is a connected object it is commonly called a train. I know how to describe a situation where they have no fusion (I did so just above) but I have no idea how to *imagine* it, or make it true in a story.

More generally, there are all sorts of puzzling sentences involving claims about **constitution**. These I think are the best guide to a solution to the puzzle.

> *A Quixotic Victory*
> –What think you of my redecorating Sancho?
> –It's rather sparse, said Sancho.
> –Sparse. Indeed it is sparse. Just a television and an armchair.
> –Where are they, Señor Quixote? asked Sancho. All I see are a knife and fork on the floor, about six feet from each other. A sparse apartment for a sparse mind. He said the last sentence under his breath so Quixote would not hear him.
> –They might look like a knife and fork, but they are a television and an armchair, replied Quixote.
> –They look just like the knife and fork I have in my pocket, said Sancho, and he moved as to put his knife and fork besides the objects on Quixote's floor.
> –Please don't do that, said Quixote, for I may be unable to tell your knife and fork from my television and armchair.
> –But if you can't tell them apart from a knife and fork, how could they be a television and an armchair?
> –Do you really think *being a television* is an observational property? asked Quixote with a grin.
> –Maybe not. OK then, how do you change the channels? asked Sancho.
> –There's a remote.
> –Where? Is it that floorboard?
> –No, it's at the repair shop, admitted Quixote.
> –I give up, said Sancho.
>
> Sancho was right to give up. Despite their odd appearance, Quixote's items of furniture really were a television and an armchair. This was the first time in months Quixote had won an argument with Sancho.

Quixote is quite right that whether something is a television is not determined entirely by how it looks. A television could be indistinguishable from a non-television. Nonetheless, something indistinguishable from a knife is not a television. Not in this world, and not in the world of *Victory* either, whatever the author says. For whether something is a television is determined at least *in part* by how it looks, and while it is impossible to provide a non-circular constraint on how a television may look, it may not look like a common knife.

In general, if whether or not something is an F is determined in part by 'lower-level' features, such as the shape and organisation of its parts, and the story specifies that the lower-level features are incompatible with the object being an F, it is not an F in the fiction. Suitably

generalised and qualified, I think this is the explanation of all of the above categories. To understand better what the generalisations and qualifications must be, we need to look at some cases that aren't like *Death*, and some alternative explanations of what is going on in *Death*.

Sentences that are **intentional errors** on the part of storytellers are not puzzling in our sense. We will use real examples for the next few pages, starting with the opening line of Joyce's most famous short story.

> *The Dead*
> Lily, the caretaker's daughter, was literally run off her feet.
> (Joyce, 1914/2000, 138)

It isn't true that Lily is *literally* run off her feet. She is run off her feet by the incoming guests, and if you asked her she may well say she was literally run off her feet, but this would reveal as much about her lack of linguistic care as about her demanding routine. Is this a case where the author loses authority over what's true in the story? No, we are not meant to read the sentence as being true in the story, but being a faithful report of what Lily (in the story) might say to herself. In practice it's incredibly difficult to tell just when the author intends a sentence to be true in the story, as opposed to being a report of some character's view of what is true. (See Holton (1997) for an illustration of the complications this can cause.) But since we are operating in theory here, we will assume *that* problem solved. The alethic puzzle only arises when it is clear that the author intends that p is true in her story, but we think p is not true. The imaginative puzzle only arises when the author invites us to imagine p, but we can not, or at least do not. Since Joyce does not intend this sentence to be true in *The Dead*, nor invites us to imagine it being true, neither puzzle arises. What happens to the phenomenological puzzle in cases like these is a little more interesting, and I'll return to that in §7.

Just as intentional errors are not puzzling, **careless errors** are not puzzling. Writing a full length novel is a perilous business. Things can go wrong. Words can be miswritten, mistyped or misprinted at several different stages. Sometimes the errors are easily detectable, sometimes they are not, especially when they concern names. In one of the drafts of *Ulysses*, Joyce managed to write "Connolly Norman" in place of "Conolly Norman". Had that draft being used for the canonical printing of the work, it would be tempting to say that we had another alethic puzzle. For the character named here is clearly the Superintendent of the Richmond District Lunatic Asylum, and his name had no double-'n', so in the story there is no double-'n' either.[1]

Here we do have an instance where what is true in the story differs from the what is written in the text. But this is not a particularly *interesting* deviation. To avoid arcane discussions of typographical errors, we will that in every case we possess an ideal version of the text, and are comparing it with the author's intentions. Slip-ups that would be detected by a careful proof-reader, whether they reveal an unintended divergence between word and world, as here,

[1] For details on the spelling of Dr Norman's name, and the story behind it, see Kidd (1988). The good doctor appears on page 6 of Joyce (1922/1993).

or between various parts of the text, as would happen if Dr Norman were not named after a real person but had his name spelled differently in parts of the text, will be ignored.[2]

Note two ways in which the puzzles as I have stated them are narrower than they first appear. First, I am only considering puzzles that arise from a particular sentence in the story, intentionally presented in the voice of an authoritative narrator. We could try and generalise, asking why it is that we sometimes (but not always) question the moral claims that are intended to be tacit in a work of fiction. For instance, we might hold that for some Shakespearean plays there are moral propositions that Shakespeare intended to be true in the play, but which are not in fact true. Such cases are interesting, but to keep the problem of manageable proportions I won't explicitly discuss them here. (I believe the solution I offer here generalises to those cases, but I won't defend that claim here.) Second, all the stories I have discussed are either paragraph-long examples, or relatively detachable parts of longer stories. For all I've said so far, the puzzle *may* be restricted to such cases. In particular, it might be the case that a suitably talented author could make it true in a story that killing people for holding up traffic is morally praiseworthy, or that a television is phenomenally and functionally indistinguishable from a knife. What we've seen so far is just that an author cannot make these things true in a story simply by saying they are true.[3] I leave open the question of whether a more subtle approach could make those things true in a fiction. Similarly, I leave it open whether a more detailed invitation to imagine that these things are true would be accepted. All we have seen so far is that simple direct invitations to imagine these things are rejected, and it feels like we could not accept them.

3 An Impossible Solution

Here's a natural solution to the puzzles, one that you may have been waiting for me to discuss. The alethic puzzle arises because only propositions that are possibly true can be true in a story, or can be imagined. The latter claim rests on the hypothesis that we can imagine only what is possible, and that we resist imagining what is impossible.

This solution assumes that it is *impossible* that killing people for holding up freeway traffic is the right thing to do. Given enough background assumptions, that *is* plausible. It is plausible, that is, that the moral facts supervene on the non-moral facts. And the supervenience principle here is quite a strong one - in *every* possible world where the descriptive facts are thus and so, the moral facts are the same way.[4] If we assume the relevant concept of impossibility is truth in no possible worlds, we get the nice result that the moral claims at the core of the problem could not possibly be true.

Several authors have discussed solutions around this area. Kendall Walton (1994) can easily be read as endorsing this solution, though Walton's discussion is rather tentative. Tamar Szabó

[2] At least, they will be ignored if it is clear they are *errors*. If there seems to be a method behind the misspellings, as in *Ulysses* there frequently is, the matter is somewhat different, and somewhat more difficult.

Tyler Doggett has argued that these cases are more similar to paradigm cases of imaginative resistance than I take them to be. Indeed, I would not have noticed the problems they raise without reading his paper. It may be a shortcoming of my theory here that I have to set questions about whether these sentences are puzzling to one side and assume an ideal proof-reader.

[3] Thanks here to George Wilson for reminding me that we haven't shown anything stronger than that.

[4] Arguably the relevant supervenience principle is even stronger than that. To use some terminology of Stephen Yablo's, there's no difference in moral facts without a difference in non-moral facts between any two *counteractual* worlds, as well as between any two *counterfactual* worlds. This might be connected to some claims I will make below about the relationship between the normative and the descriptive.

Gendler rejects the theory, but thinks it is the most natural idea, and spends much of her paper arguing against this solution. As those authors, and Gregory Currie (2002), note, the solution needs to be tidied up a little before it will work for the phenomenological and imaginative puzzles. (It is less clear whether the tidying matters to the alethic puzzle.) For one thing, there is no felt asymmetry between a story containing, "Alex proved the twin primes theorem," and one containing, "Alex found the largest pair of twin primes," even though *one of them* is impossible. Since we don't know which it is, the impossibility of the false one cannot help us here. So the theory must be that it is believed impossibilities that matter, for determining what we can imagine, not just any old impossibilities. Presumably impossibilities that are not salient will also not prevent imagination.

Even thus qualified, the solution still overgenerates, as Gendler noted. There are stories that are not puzzling in any way that contain known salient impossibilities. Gendler suggests three kinds of cases of this kind, of which I think only the third *clearly* works. The first kind of case is where we have direct contradictions true in the story. Gendler suggests that her *Tower of Goldbach* story, where seven plus five both does and does not equal twelve, is not puzzling. Graham Priest (1999) makes a similar point with a story, *Sylvan's Box*, involving an empty box with a small statue in one corner. These are clear cases of known, salient impossibility, but arguably are not puzzling in any respect. (There is a distinction between the puzzles though. It is very plausible to say that it's true in Priest's story that there's an empty box with a small statue in one corner. It is less plausible to say we really can imagine such a situation.) Opinion about such cases tends to be fairly sharply divided, and it is not good I suspect to rest too much weight on them one way or the other.

The second kind of case Gendler suggests is where we have a distinctively metaphysical impossibility, such as a singing snowman or a talking playing card. Similar cases as discussed by Alex Byrne (1993) who takes them to raise problems for David Lewis's (1978b) subjunctive conditionals account of truth in fiction. If we believe a strong enough kind of essentialism, then these will be impossible, but they clearly do not generate puzzling stories. For a quick proof of this, note that *Alice in Wonderland* is not puzzling, but several essentialist theses are violated there. It is true in *Alice in Wonderland*, for example, that playing cards plant rose trees.

But these examples don't strike me as particularly convincing either. For one thing, the essentialism assumed here may be wrong. For another, the essentialism might not be both salient and believed to be right, which is what is needed. And most importantly, we can easily reinterpret what the authors are saying in order to be make the story possibly true. We can assume, for example, that the rosebush planting playing cards are not *playing cards* as we know them, but roughly human-shaped beings with playing cards for torsos. Gendler and Byrne each say that this is to misinterpret the author, but I'm not sure this is true. As some evidence, note that the authorised illustrations in *Alice* tend to support the reinterpretations.[5]

Gendler's third case is better. There are science fiction stories, especially time travel stories, that are clearly impossible but which do not generate resistance. Here's two such stories, the

[5]Determining whether this is true in *all* such stories would be an enormous task, I fear, and somewhat pointless given the next objection. If anyone wants to say all clearly impossible statements in fiction are puzzling, I suspect the best strategy is to divide and conquer. The most blatantly impossible claims are most naturally fit for reinterpretation, and the other claims rest on an essentialism that is arguably not proven. I won't try such a massive defence of a false theory *here*.

first lightly modified from a surprisingly popular movie, and the second lifted straight from a very popular source.

> *Back to the Future*
> Marty McFly unintentionally travelled back in time to escape some marauding Libyan terrorists. In doing so he prevented the chance meeting which had, in the timeline that had been, caused his father and mother to start dating. Without that event, his mother saw no reason to date the unattractive, boring nerdy kid who had been, in a history that no longer is, Marty's father. So Marty never came into existence. This was really a neat trick on Marty's part, though he was of course no longer around to appreciate it. Some people manage to remove themselves from the future of the world by foolish actions involving cars. Marty managed to remove himself from the past as well.

> *The Restaurant at the End of the Universe*
> The Restaurant at the End of the Universe is one of the most extraordinary ventures in the entire history of catering.
>
> It is built on the fragmented remains of an eventually ruined planet which is enclosed in a vast time bubble and projected forward in time to the precise moment of the End of the Universe.
>
> This is, many would say, impossible.
>
> …
>
> You can visit it as many times as you like … and be sure of never meeting yourself, because of the embarrassment this usually causes.
>
> This, even if the rest were true, which it isn't, is patently impossible, say the doubters.
>
> All you have to do is deposit one penny in a savings account in your own era, and when you arrive at the End of Time the operation of compound interest means that the fabulous cost of your meal has been paid for.
>
> This, many claim, is not merely impossible but clearly insane. (Adams, 1980, 213-214)

Neither of these are puzzling. Perhaps it's hard to imagine the last couple of sentences of the McFly story, but everything the respective authors say is true in their stories. So the impossibility theory cannot be right, because it overgenerates, just as Gendler said.

Recently Kathleen Stock (2003) has argued that one of the assumptions that Gendler makes, specifically that it isn't true that "a judgement of conceptual impossibility renders a scenario unimaginable" (Gendler, 2000, 66) is false. Even if Stock is right, this doesn't threaten the kind of response that I have (following Gendler) offered to the puzzles. But actually there are a few reasons to doubt Stock's reply. I'll discuss these points in order.

It isn't entirely clear from Stock's discussion what she is taking a conceptual impossibility to be. I *think* it is a proposition of the form *Some F is a G* (or *That F is a G*, or something of this

sort) where it is constitutive of being an F that the F is not a G. There is no positive characterisation of conceptual impossibility in Stock's paper, but it is clearly meant to be something stronger than mere impossibility, or a priori falsehood. In any case, most of the core arguments turn on worries about allegedly deploying a concept while refusing to draw inferences that are constitutive of that concept, so the kind of definition I've offered above seems to be on the right track.

Now if this is the case then Stock has no objection to the imaginability of the two stories I offered that involve known and salient impossibilities. For neither of these stories includes a conceptual impossibility in this sense. So even if conceptual impossibilities cannot be imagined, some impossibilities can be imagined. (And at this point what holds for imagination also holds for truth in fiction.)

While this suffices as a response to the particular claims Stock makes, it might be thought it undercuts the objection I have made to the impossible solution. For it might be thought that what is wrong with the puzzling sentences just is that they represent *conceptual* impossibilities in this sense, and we have no argument that these can be imagined, or true in fiction. This is not too far removed from the actual solution I will offer, so it is a serious worry. The problem with this line is that not all of our puzzles are conceptual impossibilities. It isn't constitutive of being a television that a thing is phenomenally or functionally distinguishable from a knife, but the claim in *Victory* that some television is not phenomenally or functionally distinguishable from a knife is puzzling. Even in our core cases, of morally deviant claims in fiction, there need not be any conceptual impossibilities. As R. M. Hare (1951) pointed out long ago, people with very different moral beliefs could have in common the concept GOOD. Arguably, someone who thinks that what Craig does in *Death* is good is morally confused, not conceptually confused. So whether Gendler or Stock is right about the imaginability of conceptual impossibility is neither here nor there with respect to these puzzles.

Having said that, there are some reasons to doubt Stock's argument. One of her moves is to argue that we couldn't imagine conceptual impossibilities because we can't believe conceptual impossibilities. But as Sorensen (2001) persuasively argues, we *can* believe conceptual impossibilities. One of Sorensen's arguments, lightly modified, helps us respond to another of Stock's arguments. Stock notes, rightly, that we shouldn't take the fact that it *seems* we can imagine impossibilities to be conclusive evidence we can do so. After all, we are wrong about whether things are as they seem all the time. But this might be a special case. I think that if it seems to be the case that p then we can imagine that p. And Stock agrees it *seems* to be the case that we can imagine conceptual impossibilities. So we can imagine that we can imagine conceptual impossibilities. Hence it can't be a conceptual impossibility that we can imagine at least one conceptual impossibility. This doesn't tell against the claim that it is some other kind of impossibility, though as we'll see Stock's main argument rests on considerations about the conceptual structure of imagination, so it isn't clear how she could argue for this.

The main argument Stock offers is that no account of how concepts work are compatible with our imagining conceptual impossibilities. Her argument that atomist theories of concepts (as in Fodor (1998)) are incompatible with imagining conceptual impossibilities isn't that persuasive. She writes that "clearly it is not the case that imagining "the cow jumped over the moon" stands in a lawful relation to the property of being a cow (let alone the property of [being] a cow jumping over the moon. Imagining by its very nature is resistant to any attempt to incorporate it into an externalist theory of content" (2003, 114). But this isn't clear at all.

When I imagine going out drinking with Bill Clinton there is, indeed there must be, some kind of causal chain running back from my imagining to Bill Clinton himself. If there was not, I'd at most be imagining going out drinking with a guy who looks a lot like Bill Clinton. Perhaps it isn't as clear, but when I imagine that a cow (and not just a zebra disguised to look like a cow) is jumping over the moon it's nomologically necessary that there's a causal chain of the right kind stretching back to actual cows. And it's arguable that the concept I deploy in imagining that a cow (a real *cow*) is jumping over the moon just is the concept whose content is fixed by the lawful connections between various cows and my (initial) deployment of it. So I don't see why a conceptual atomist should find this kind of argument convincing.

Stock's response to Gendler was presented at a conference on Imagination and the Arts at Leeds in 2001, and at the same conference Derek Matravers (2003) offered an alternative solution to the alethic puzzle. Although it does not rest on claims about impossibility, it also suffers from an overgeneration problem. Matravers suggests that in at least some fictions, we treat the text as a report by a (fictional) narrator concerning what is going on in a faraway land. Now in reality when we hear reports from generally trustworthy foreign correspondents, we are likely to believe their descriptive claims about the facts on the ground. Since they have travelled to the lands in question, and we have not, the correspondent is epistemologically privileged with respect to those facts on the ground. But when the correspondent makes moral evaluations of those facts, she is not in a privileged position, so we don't just take her claims as the final word. Matravers suggests there are analogous limits to how far we trust a fictional narrator.

The problem with this approach is that there are several salient disanalogies between the position of the correspondent and the fictional narrator. The following case, which I heard about from Mark Liberman, illustrates this nicely. On March 5, 2004, the BBC reported that children in a nursery in England had found a frog with three heads and six legs. Many people, including Professor Liberman, were sceptical, notwithstanding the fact that the BBC was actually in England and Professor Liberman was not. The epistemological privilege generated by proximity doesn't extend to implausible claims about three-headed frogs. The obvious disanalogy is that if a fictional narrator said that there was a three-headed six-legged frog in the children's nursery then other things being equal we would infer it is true in the fiction that there was indeed a three-headed six-legged frog in the children's nursery.[6] So there isn't an easy analogy between when we trust foreign correspondents and fictional narrators. Now we need an explanation of why the analogy does hold when either party makes morally deviant claims, even though it doesn't when they both make biologically deviant claims. But it doesn't seem any easier to say why the analogy holds then than it is to solve the original puzzle.

Two other quick points about Matravers's solution. It's going to be a little delicate to extend this solution to all the cases I have discussed above, for normally we do think fictional narrators are privileged with respect to where the televisions and windows are. What matters here is that how far narratorial privilege extends depends on what other claims the narrator makes. Perhaps the same is true of foreign correspondents, though we'd need to see an argument for that. Second, it isn't clear how this solution could possibly generalise to cover cases, such as frequently occurs in plays, where the deviant moral claim is clearly intended by the author to be

[6]There is a complication here in that such a sentence *might* be evidence that the fictional work is not to be understood as this kind of report, and instead understood as something like a recording of the children's thoughts. I'll assume we're in a story where it is clear that the sentences are not to be so interpreted.

true in the fiction but the reader (or watcher) does not agree even though the author's intention is recognised. As I mentioned at the start, these cases aren't our concern here, though it would be nice to see how a generalisation to these cases is possible. But the primary problem with Matravers's solution is that as it stands it (improperly) rules out three-headed frogs in fiction, and it is hard to see how to remedy this problem without solving the original puzzle.

4 Some Ethical Solutions

If one focuses on cases like *Death*, it is natural to think the puzzle probably has something to do with the special nature of ethical predicates, or perhaps of ethical concepts, or perhaps of the role of either of these in fiction. I don't think any such solution can work because it can't explain what goes wrong in *Victory*, and this will recur as an objection in what follows.

The most detailed solution to the puzzles has been put forward by Tamar Szabó Gendler. She focuses on the imaginative puzzle, but she also makes valuable points about the other puzzles. My solution to the phenomenological puzzle is basically hers plus a little epicycle.

She says that we do not imagine morally deviant fictional worlds because of our "general desire to not be manipulated into taking on points of view that we would not reflectively endorse as our own." How could we take on a point of view by accepting something in a fiction? Because of the phenomena noted above that some things become true in a story *because* they are true in the world. If this is right, its converse must be true as well. If what is true in the story must match what is true in the world, then to accept that something is true in the story just is to accept that it is true in the world. Arguably, the same kind of 'import/export' principles hold for imagination as for truth in fiction. Some propositions become part of the content of an imagining because they are true. So, in the right circumstances, they will only be part of an imagining if they are true. Hence to imagine them (in the right circumstances) is to commit oneself to their truth. Gendler holds that we are sensitive to this phenomena, and that we refuse to accept stories that are morally deviant because that would involve accepting that morally deviant claims are true in the world.

That's a relatively rough description of Gendler's theory, but it says enough to illustrate what she has in mind, and to show where two objections may slip in. First, it is not clear that it generalises to all the cases. Gendler is aware of some of these cases and just bites the relevant bullets. She holds, for instance, that we can imagine that actually lame jokes are funny, and it could be true in a story that such a joke is funny. It would be a serious cost to her theory if she had to say the same thing about *all* the examples discussed above.

The second problem is more serious. The solution is only as good as the claim that moral claims are more easily exported than descriptive claims, and more generally that the types of claims we won't imagine are more easily exported than those we don't resist. Gendler has two arguments for why the first of these should be true, but neither of them sounds persuasive. First, she says that the moral claims are true in all possible worlds if true at all. But this won't do on its own, because *as she proved*, we don't resist some necessarily false claims. (This objection is also made by (Matravers, 2003, 94).)

Secondly, she claims that in other cases where there are necessary falsehoods true in a story, as in *Alice in Wonderland*, or the science fiction cases, the author makes it clear that unusual export restrictions are being imposed. But this is wrong for two reasons. First, I don't think that any particularly clear signal to this effect occurs in my version of *Back to the Future*. Secondly,

even if I had explicitly signalled that I had intended to make some of the facts in the story available for export, and you didn't believe that, that isn't enough reason to resist imagining the story. For my intent as to what can and cannot be exported is not part of the story.

To see this, consider one relatively famous example. At one stage B. F. Skinner tried to promote behaviourism by weaving his theories into a novel (of sorts): *Walden Two*. Now I'm sure Skinner intended us to export some psychological and political claims from the story to the real world. But it is entirely possible to read the story with full export restrictions in force without rejecting that what Skinner says is true in that world. (It is dreadfully boring, since there's nothing but propagandising going on, but *possible*.) If exporting was the only barrier here, we should be able to impose our own tariff walls and read the story along, whatever the intent of the author, as we can with *Walden Two*. One can accept it is true in *Walden Two* that behaviourism is the basis of a successful social policy, even though Skinner wants us to accept this as true in the story iff it is true in the world, and it isn't true in the world. We cannot read *Death* or *Victory* with the same ironic detachment, and Gendler's theory lacks the resources to explain this.

Currie's theory attacks the problem from a quite different direction. He relies on the motivational consequences of accepting moral claims. Assume internalism about moral motivation, so to accept that ϕ-ing is right is to be motivated to ϕ, at least ceteris paribus. So accepting that ϕ-ing is right involves acquiring a desire to ϕ, as well, perhaps, as beliefs about ϕ-ing. Currie suggests that there is a mental state that stands to desire the way that ordinary imagination stands to belief. It is, roughly, a state of having an off-line desire, in the way that imagining that p is like having an off-line belief that p, a state like a belief that p but without the motivational consequences. Currie suggests that imagining that ϕ-ing is right involves off-line acceptance that ϕ-ing is right, and that in part involves having an off-line desire (a desire-like imagination) to ϕ. Finally, Currie says, it is harder to alter our off-line desires at will than it is to alter our off-line beliefs, and this explains the asymmetry. The argument for this last claim seems very hasty, but we'll let that pass. For even if it is true, Currie's theory does little to explain the later cases of imaginative resistance, from *Alien Robbery* to *Victory*. It cannot explain, why we have resistance to claims about what is rational to believe, or what is beautiful, or what attitudes other people have. The idea that there is a state that stands to desire as imagination stands to belief is I suspect a very fruitful one, but I don't think its fruits include a solution to these puzzles.

5 Grok

Stephen Yablo has suggested that the puzzles, or at least the imaginative puzzle, is closely linked to what he calls *response-enabled* concepts, or *grokking* concepts. (I'll also use response-enabled (grokking) as a property of the predicates that pick out these concepts.) These are introduced by examples, particularly by the example 'oval'.

Here are meant to be some platitudes about OVAL. It is a shape concept - any two objects in any two worlds, counterfactual or counteractual, that have the same shape are alike in whether they are ovals. But which shape concept it is is picked out by our reactions. They are the shapes that strike us as being egg-like, or perhaps more formally, like the shape of all ellipses whose length/width ratio is the golden ratio. In this way the concept OVAL meant to be distinguished on the one hand from, say, PRIME NUMBER, which is entirely independent

of us, and from WATER, which would have picked out a different chemical substance had our reactions to various chemicals been different. Note that what 'prime number' picks out is determined by us, like all semantic facts are. So the move space into which OVAL is meant to fit is quite tiny. We matter to its extension, but not the way we matter to 'prime number' (or that we don't matter to PRIME NUMBER), and not the way we matter to 'water'. I'm not sure there's any space here at all. To my ear, Yablo's grokking predicates strike me as words that have associated egocentric descriptions that fix their reference without having egocentric reference fixing descriptions, and such words presumably don't exist. But for present purposes I'll bracket those general concerns and see how this idea can help solve the puzzles. For despite my disagreement about what these puzzles show about the theory of concepts, Yablo's solution is not too dissimilar to mine.

The important point for fiction about grokking concepts is that we matter, in a non-constitutive way, for their extension. Not we as we might have been, or we as we are in a story, but us. So an author can't say that in the story squares looked egg-shaped to the people, so in the story squares are ovals, because *we* get to say what's an oval, not some fictional character. Here's how Yablo puts it:

> Why should resistance [meaning, roughly, unimaginability] and grokkingness be connected in this way? It's a feature of grokking concepts that their extension in a situation depends on how the situation does or would strike us. 'Does or would strike us' *as we are*: how we are represented as reacting, or invited to react, has nothing to do with it. Resistance is the natural consequence. If we insist on judging the extension ourselves, it stands to reason that any seeming intelligence coming from elsewhere is automatically suspect. This applies in particular to being 'told' about the extension by an as-if knowledgeable narrator. (2002, 485)

It might look at first as if *Victory* will be a counterexample to Yablo's solution, just as it is to the Ethical solutions. After all, the concept that seems to generate the puzzles there is TELEVISION, and that isn't at all like his examples of grokking concepts. (The examples, apart from evaluative concepts, are all shape concepts.) On the other hand, if there are any grokking concepts, perhaps it is plausible that TELEVISION should be one of them. Indeed, the platitudes about TELEVISION provide some support for this. (The following two paragraphs rely heavily on Fodor (1998).)

Three platitudes about TELEVISION stand out. One is that it's very hard to define just what a television is. A second is that there's a striking correlation between people who have the concept TELEVISION and people who have been acquainted with a television. Not a perfect correlation - some infants have acquaintance with televisions but not as such, and some people acquire TELEVISION by description - but still strikingly high. And a third is that conversations about televisions are rarely at cross purposes, even when they consist of people literally talking different languages. TELEVISION is a shared concept.

Can we put these into a theory of the concept TELEVISION? Fodor suggests we can, as long as we are not looking for an analysis of TELEVISION. Televisions are those things that strike us, people in general, as being sufficiently like the televisions we've seen, in a televisual kind of way. This isn't an account of the meaning of the word 'television' - there's no reference to us in that word's dictionary entry, and rightly so. Nor is it an analysis of what constitutes

the concept television. There's no reference to us there either. But it does latch on to the right concept, or at least the right extension, in perhaps the only way we could. And this proposal certainly explains the platitudes well. The epistemic necessity of having a paradigm television to use as a basis for similarity judgments explains the striking correlation between televisual acquaintance and concept possession. The fact that the only way of picking out the extension uses something that is not constitutive of the concept, namely our reactions to televisions, explains why we can't reductively analyse the concept. And the use of people's reactions in general rather than idiosyncratic reactions explains why its a common concept. These look like good reasons to think something like Fodor's theory of the concept TELEVISION is right, and if it is then TELEVISION seems to be response-enabled in Yablo's sense. So unlike the Ethical solutions, Yablo's solution might yet predict that *Victory* will be puzzling.

Still, I have three quibbles about his solution, and that's enough to make me think a better solution may still to be found.

First, there's a missing antecedent in a key sentence in his account, and it's hard to see how to fill it in. What does he mean when he says 'how the situation does or would strike us'? Does or would strike us if *what*? If we were there? But we don't know where there is. There, in *Victory*, is allegedly a place where televisions look like knifes and forks. What if the antecedent is *If all the non-grokking descriptions were accurate*? The problem now is that this will be too light. If TELEVISION is grokking, then there is a worry that many concepts, including perhaps all artefact concepts, will be grokking. Fodor didn't illustrate his theory with TELEVISION, he always used DOORKNOB. But the theory was meant to be rather general. If we take out all the claims involving grokking concepts, there may not be much left.

Second, despite the generality of Fodor's account, it isn't clear that mental concepts, and content concepts, are grokking. We would need another argument that LOVE is grokking, and that so is BELIEVING THAT THERE ARE SPACE ALIENS. Perhaps such an argument can be given, but it will not be a trivial exercise.

Finally, I think this Yablo's solution, at least as most naturally interpreted, over-generalises. Here's a counterexample to it. The following story is not, I take it, puzzling.

> *Fixing a Hole*
> DQ and his buddy SP leave DQ's apartment at midday Tuesday, leaving a well-arranged lounge suite and home theatre unit, featuring DQ's prized oval television. They travel back in time to Monday, where DQ has some rather strange and unexpected adventures. He intended to correct something that happened yesterday, that had gone all wrong the first time around, and by the time the buddies reunite and leave for Tuesday (by sleeping and waking up in the future) he's sure it's all been sorted. When DQ and his buddy SP get back to his apartment midday Tuesday, it looks for all the world like there's nothing there except an ordinary knife and fork.

Now this situation would not strike us, were we to see it, as one where there is a lounge suite and home theatre unit in DQ's apartment midday Tuesday, for it looks as if there's an ordinary knife and fork there. But still, the author gets to say that what's in DQ's apartment as the story opens includes an oval television. And this despite the fact that the two concepts, TELEVISION and OVAL, are grokking. Perhaps some epicycles could be added to Yablo's theory to solve this problem, but for now the solution is incomplete.

6 Virtue

The content cases may remind us of one of Fodor's most famous lines about meaning.

> I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin*, *charm*, and *charge* will perhaps appear on the list. But *aboutness* surely won't; intentionality doesn't go that deep … If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe their supervenience on?) properties that are themselves neither intentional nor semantic. If aboutness is real, it must really be something else. (Fodor, 1987, 97)

If meaning doesn't go that deep, but there are meaning facts, then those facts must hold in virtue of more fundamental facts. "Molino de viento" means windmill in Spanish in virtue of a pattern of usage of those words by Spanish speakers, for instance.

It seems that many of the stories above involve facts that hold, if they hold at all, in virtue of other facts. Had Fodor other interests than intentionality, he may have written instead that beauty doesn't go that deep, and neither does television. If an event is to be beautiful, this is a fact that must obtain in virtue of other facts about it, perhaps its integrity, wholeness, symmetry and radiance as Aquinas says (Joyce, 1944/1963, 212), and that event being a monster truck death match of doom *probably* precludes those facts from obtaining.[7] If Quixote's favourite item of furniture is to be a television, this must be in virtue of it filling certain functional roles, and being indistinguishable from a common knife probably precludes that.

What is it for a fact to obtain in virtue of other facts obtaining? A good question, but not one we will answer here. Still, the concept seems clear enough that we can still use it, as Fodor does. What we have in mind by 'virtue' is understandable from the examples. One thing to note from the top is that it is not just supervenience: whether x is good supervenes on whether it is good, but it is not good *in virtue of* being good. How much our concept differs from supervenience is a little delicate, but it certainly differs.

Returning to our original example, moral properties are also less than perfectly fundamental. It is not a primitive fact that the butcher or the baker is generous, but a fact that obtains in virtue of the way they treat their neighbours. It is not a primitive fact that what Craig does is wrong, but a fact that obtains in virtue of the physical features of his actions.

How are these virtuous relations relevant to the puzzles? To a first approximation, these relations are always imported into stories and into imagination. The puzzles arise when we try to tell stories or imagine scenes where they are violated. The rest of the paper will be concerned with making this claim more precise, motivating it, and arguing that it solves the puzzles. In making the claim precise, we will largely be qualifying it.

The first qualification follows from something we noted at the end of section 2. We don't know whether puzzles like the ones with which we started arise whenever there is a clash between real-world morality (or epistemology or mereology) and the morality (or epistemology or mereology) the author tries to put in the story. We do know they arise for simple stories and direct invitations to imagine. So if we aren't to make claims that go beyond our evidence, we

[7] Although it isn't obvious just which of the Thomistic properties the death match lacks.

should say there is a default assumption that these relations are imported into stories or imaginations, and it is not easy to overcome this assumption. (I will say for short there is a strong default assumption, meaning just that an author cannot cancel the assumption by saying so, and that we cannot easily follow invitations to imagine that violate the relations.)

The second qualification is that sometimes we simply ignore, either in fiction or imagination, what goes on at some levels of detail. This means that sometimes, in a sense, the relations are not imported into the story. For instance, for it to really be true that in a language that "glory" means *a nice knockdown argument*, this must be true in virtue of facts about how the speakers of that language use, or are disposed to use, "glory". But we can simply say in a story that "glory" in a character's language means *a nice knockdown argument* without thereby making any more general facts about usage or disposition to use true in the story.[8] More generally, we can simply pick a level of conceptual complexity at which to write our story or conduct our imaginings. Even if those concepts apply, when they do, in virtue of more basic facts, no more basic facts need be imported into the story. For a more vivid, if more controversial, example, one might think that cows are cows in virtue of their DNA having certain chemical characteristics. But when we imagine a cow jumping over the moon, we need not imagine anything about chemistry. Those facts are simply below the radar of our imagining. What do we mean then when we say that these relations are imported into the story? Just that if the story regards both the higher-level facts and the lower-level facts as being within its purview, then they must match up. This does not rule out the possibility of simply leaving out all lower-level facts from the story. In general the same thing is true for imagining, though we will look at some cases below where we it seems there is a stronger constraint on imagining.

The third qualification is needed to handle an example pressed on me by a referee. Recall our example *Fixing a Hole*.

> *Fixing a Hole*
> DQ and his buddy SP leave DQ's apartment at midday Tuesday, leaving a well-arranged lounge suite and home theatre unit, featuring DQ's prized oval television. They travel back in time to Monday, where DQ has some rather strange and unexpected adventures. He intended to correct something that happened yesterday, that had gone all wrong the first time around, and by the time the buddies reunite and leave for Tuesday (by sleeping and waking up in the future) he's sure it's all been sorted. When DQ and his buddy SP get back to his apartment midday Tuesday, it looks for all the world like there's nothing there except an ordinary knife and fork.

In this story it seems that on Tuesday there is a television that looks exactly like a knife. If we interpret the claim about the relations between higher-level facts and the lower-level facts as a kind of impossibility claim, e.g. as the claim that a conjunction $p \wedge q$ is never true in a story if the conditional *If q, then p is false in virtue of q being true* is true, then we have a problem. Let p be the claim that there is a television, and let q be the claim that the only things in the apartment looked life a knife and fork. If that's how the more basic phenomenal and functional facts are,

[8] Do we make facts about the actual speaker's usage true in the story? No. The character might have idiosyncratic reasons for not using the word "glory", and for ignoring all others who use it. That's consistent with the word meaning a nice knockdown argument.

then there isn't a television in virtue of those facts. (That is, this relation between phenomenal and functional facts and facts about where the televisions are really holds.) So this rule would say $p \wedge q$ could not be true in the story. But in fact $p \wedge q$ is true in the story.

The difficulty here is that *Fixing a Hole* is a contradictory story, and contradictory stories need care. First, here's how we should interpret the rule

> *Virtue*
> If p is the kind of claim that if true must be true in virtue of lower-level facts, and if the story is about those lower-level facts, then it must be true in the story that there is some true proposition r which is about those lower-level facts such that p is true in virtue of r.

In *Fixing a Hole* there are some true lower-level claims that are inconsistent with there being a television. But there is also in the story a true proposition about how DQ's television looked before his time-travel misadventure. And it is true (both in reality and in the story) that something is a television in virtue of looking that way. (Note that we don't say there must be some proposition r that is true in the story in virtue of which p is true. For there is no fact of the matter in *Fixing a Hole* about how DQ's television looked before he left. So in reality we could not find such a proposition. But it is true in the story that his television looks some way or other, so as long as we talk about what in the story is true, and don't quantify over propositions that are (in reality) true in the story, we avoid this pitfall.)

So my solution to the alethic puzzle is that *Virtue* is a strong default principle of fictional interpretation. I haven't done much yet to motivate it, apart from noting that it seems to cover a lot of the cases that have been raised without overgenerating in the manner of the impossible solution. A more positive motivation must wait until I have presented my solutions to the phenomenological and imaginative puzzles. I'll do that in the next section, then in §8 tell a story about why we should believe *Virtue*.

7 More Solutions

7.1 The Phenomenological Puzzle

My solution here is essentially the same as Gendler's. She think that when we strike a sentence that generated imaginative resistance we respond with something like, "That's what you think!" What makes this notable is that it's constitutive of playing the fiction game that we not normally respond that we way, that we give the author some flexibility in setting up a world. I think that's basically right, but a little more is needed to put the puzzle to bed.

Sometimes the "That's what you think!" response does not constitute abandoning the fiction game. At times it is the only correct way to play the game. It's the right thing to say *to Lily* when reading the first line of *The Dead*. (Maybe it would be rude to say it *aloud* to poor Lily, the poor girl is run off her feet after all, but it's appropriate to think it.) This pattern recurs throughout *Dubliners*. When in *Eveline* the narrator says that Frank has sailed around the world, the right reaction is to say to Eveline (or whoever is narrating then), "That's what you think!" There's a cost to playing the game this way. We end up knowing next to nothing about Frank. But it is not as if making the move stops us playing, or even stops us playing correctly. It's part of the point of *Eveline* that we know next to nothing about Frank.

What makes cases like *Death* and *Victory* odd is that our reaction is directed at someone who isn't in the story. One of Alex Byrne's (1993) criticisms of Lewis was that on Lewis's theory it is true in every story that the story is being told. Byrne argued that in many fictions it is not true that in the fictional world there is someone sufficiently knowledgeable to tell the story. In these fictions, we have a story without a storyteller. If there are such stories, then presumably *Death* and *Victory* are amongst them. It is not a character in the story who ends by saying that Craig's action was right or that Quixote's apartment contains a television. The *author* says that, and hence deserves our reproach, but the author isn't in the story. Saying "That's what you think!" directly to him or her breaks the fictional spell for suddenly we have to recognise a character not in the fictional world.

This proposal for the phenomenological puzzle yields a number of predictions which seem to be true and interesting. First, a story that has a narrator should not generate a phenomenological puzzle, even when outlandish moral claims are made. The more prominent the narrator, the less striking the moral claim. Imagine, for example, a version of *Death* where the text purports to be Craig's diary, and it includes naturally enough his own positive evaluation of what he did. We wouldn't believe him, of course, but we wouldn't be struck by the claim the same way we are in the actual version of *Death*.

One might have thought that what is shocking is what we discover about the author. But this isn't right, as can be seen if we reflect on stories that contain Craig's diary. It is possible, difficult but possible, to embed the diary entry corresponding to *Death* in a longer story where it is clear that the author endorses Craig's opinions. (Naturally I won't do this. Examples have to come to an end somewhere.) Such a story would, in a way, be incredibly shocking. But it wouldn't make the final line shocking in just the way that the final line of *Death* is shocking. Our reactions to these cases suggest that the strikingness of the last line of *Death* is not a function of what it reveals about the author, but of how it reveals it.

The final prediction my theory makes is somewhat more contentious. Some novels announce themselves as works of fiction. They go out of their way to prevent you ignoring the novel's role as mediation to a fictional world. (For an early example of this, consider the sudden appearance of newspaper headlines in the 'Aeolus' episode of *Ulysses*.) In such novels we already have to recognise the author as a player in the fictional game, if not a character in the story. I predict that sentences where we do not take what is written to really be true in the story, even though this is what the author intended, should be less striking in these cases because we are already used to reacting to the author *as such* rather than just to the characters. Such books go out of their way to break the fictional spell, so spell breaking should matter less in these cases. I think this prediction is correct, although the works in question tend to be so complicated that it is hard to generate clear intuitions about them.

7.2 The Imaginative Puzzle

Imagine, if you will, a chair. Have you done so? Good. Let me make some guesses about what you imagined. First, it was a specific kind of chair. There is a fact of the matter about whether the chair you imagined is, for example, an armchair or a dining chair or a classroom chair or an airport lounge chair or an outdoor chair or an electric chair or a throne. We can verbally

represent something as being a chair without representing it as being a specific kind of chair, but imagination cannot be quite so coarse.[9]

Secondly, what you imagined was incomplete in some respects. You possibly imagined a chair that if realised would contain some stitching somewhere, but you did not imagine any details about the stitching. There is no fact of the matter about how the chair you imagined holds together, if indeed it does. If you imagined a chair by imagining bumping into something chair-like in the dead of night, you need not have imagined a chair of any colour, although in reality the chair would have some colour or other.[10]

Were my guesses correct? Good. The little I needed to know about imagination to get those guesses right goes a long way towards solving the puzzle.

Chairs are not very distinctive. Whenever we try to imagine that a non-fundamental property is instantiated the content of our imagining will be to some extent more specific than just that the object imagined has the property, but not so much more specific as to amount to a complete description of a possibilia. It's the latter fact that does the work in explaining how can imagine impossible situations. If we were, foolishly, to try to fill in all the details of the impossible science fiction cases it would be clear they contained not just impossibilities, but violations of *Virtue*, and then we would no longer be able to imagine them. But we can imagine the restaurant at the end of the universe without imagining it in all its glorious gory detail. And when we do so our imagining appears to contain no such violations.

But why can't we imagine these violations in fictions? It is primarily because we can only imagine the higher-level claim some way or another, just as we only imagine a chair as some chair or other, and the instructions that go along with the fiction forbid us from imagining *any* relevant lower-level facts that would constitute the truth of the higher-level claim. We have not stressed it much above, but it is relevant that fictions understood as invitations to imagine have a "That's all" clause.[11] We are not imagining *Death* if we imagine that Jack and Jill had just stopped arguing with each other and were about to shoot everyone in sight when Craig shot them in self-defence. The story does not explicitly say that wasn't about to happen. It doesn't include a "That's all" clause. But such clauses have to be understood. So not only are we instructed to imagine something that seems incompatible with Craig's action being morally acceptable, we are also instructed (tacitly) to *not* imagine anything that would make it the case that his action is morally acceptable. But we can't simply imagine moral goodness in the abstract, to imagine it we have to imagine a particular kind of goodness.

---

[9] This relates to another area in which my solution owes a debt to Gendler's solution. Supposing can be coarse in a way that imagining cannot. We can suppose that Jack sold a chair without supposing that he sold an armchair or a dining chair or any particular kind of chair at all. Gendler concludes that what we do in fiction, where we try and imagine the fictional world, is very different to what we do, say, in philosophical argumentation, where we often suppose that things are different to the way they actually are. We can suppose, for the sake of argument as it's put, that Kantian or Aristotelian ethical theories are entirely correct, even if we have no idea how to imagine either being correct. Thanks to Tyler Doggett for pointing out the connection to Gendler here.

[10] Thanks to Kendall Walton for pointing out this possibility.

[11] "That's all" clauses play a distinct, but related, role in (Jackson, 1998, Ch. 1). It's also crucial to my solution to the alethic puzzle that there be a "That's all" clause in the story. What's problematic about these cases is that the story (implicitly) rules out there being the lower-level facts that would make the expressed higher-level claims true.

7.3 Two Thoughts Too Many?

I have presented three solutions to the three different puzzles with which we started. Might it not be better to have a uniform solution? No, because although the puzzles are related, they are not identical. Three puzzles demand three solutions.

We saw already that the phenomenological puzzle is different to the other two. If we rewrite *Death* as Craig's diary there would be nothing particularly striking about the last sentence, certainly in the context of the story as so told. But the last sentence generates alethic and imaginative puzzles. Or at least it could generate these puzzles if the author has made it clear elsewhere in the story that Craig's voice is authoritative. So we shouldn't expect the same solution to that puzzle as the other two.

The alethic puzzle is different to the other two because ultimately it depends on what the moral and conceptual truths are not on what we take them to be. Consider the following story.

> *The Benefactor*
> Smith was a very generous, just and in every respect moral man. Every month he held a giant feast for the village where they were able to escape their usual diet of gains, fruits and vegetables to eat the many and varied meats that Smith provided for them.

Consider in particular, what should be easy to some, how *Benefactor* reads to someone who believes that we are morally required to be vegetarian if this is feasible. In *Benefactor* it is clear in the story that most villagers can survive on a vegetarian diet. So it is morally wrong to serve them the many and varied meats that Smith does. Hence such a reader should disagree with the author's assessment that Smith is moral 'in every respect'. Such a reader will think that in fact in the story Smith is quite immoral in one important respect.

Now for our final assumption. Assume it is really true that we morally shouldn't eat meat if it is avoidable. Since the ethical vegetarians have true ethical beliefs about the salient facts here, it seems plausible that their views on what is true in the story should carry more weight than ours. (I'm just relying on a general epistemological principle here: other things being equal trust the people who have true beliefs about the relevant background facts.) So it seems that it really is false in the story that Smith is in every respect moral. *Benefactor* raises an alethic puzzle even though for non-vegetarians it does not raise a phenomenological or imaginative puzzle.

This point generalises, so we need not assume for the general point that vegetarianism is true or that our typical reader is not vegetarian. We can be very confident that *some* of our ethical views will be wrong, though for obvious reasons it is hard to say which ones. Let p be a false moral belief that we have. And let S be a story in which p is asserted by the (would-be omniscient) narrator. For reasons similar to what we said about *Benefactor*, p is not true in S. But S need not raise any imaginative or phenomenological puzzles. Hence the alethic puzzle is different to the other two puzzles.

8 Why Virtue Matters

I owe you an argument for why authors should be unable to easily generate violations of *Virtue*, though there is no general bar on making impossibilities true in a story. My general claims here are not too dissimilar to Yablo's solution to the puzzles, but there are a couple of distinctive new points. Before we get to the argument, it's time for another story.

Three design students walk into an furniture showroom. The new season's fashions are all on display. The students are all struck by the *piece de resistance*, though they are all differently struck by it. Over drinks later, it is revealed that while B and C thought it was a chair, A did not. But the differences did not end there. When asked to sketch this contentious object, A and B produced identical sketches, while C's recollections were drawn somewhat differently. B clearly disagrees with both A and C, but her differences with each are quite different. With C she disagrees on some simple empirical facts, what the object in question looked like. With A she disagrees on a conceptual fact, or perhaps a semantic fact, whether the concept CHAIR, or perhaps just the term 'chair', applies to the object in question. As it turns out, A and B agree that 'chair' means CHAIR, and agree that CHAIR is a public concept so one of them is right and the other wrong about whether this object falls under the concept. In this case, their disagreement will have a quite different feel to B's disagreement with C. It may well be that there is no analytic/synthetic distinction, and that questions about whether an object satisfies a concept are always empirical questions, but this is not how it feels to A and B. They feel that they agree on what the world is like, or at least what this significant portion of it is like, and disagree just on which concepts apply to it.

The difference between these two kinds of disagreement is at the basis of our attitudes towards the alethic puzzle. It may look like we are severely cramping authorial freedom by not permitting violations of *Virtue*.[12] From A and B's perspective, however, this is no restriction at all. Authors, they think, are free to stipulate which world will be the site of their fiction. But as their disagreement about whether the *piece de resistance* was a chair showed, we can agree about which world we are discussing and disagree about which concepts apply to it. The important point is that the metaphysics and epistemology of concepts comes apart here.

There can be no difference in whether the concept CHAIR applies without a difference in the underlying facts. But there can be a difference of opinion about whether a thing is a chair without a difference of opinion about the underlying facts. The fact that it's the author's story, not the reader's, means that the author gets to say what the underlying facts are. But that still leaves the possibility for differences of opinion about whether there are chairs, and on that question the author's opinion is just another opinion.

Authorial authority extends as far as saying which world is fictional in their story, it does not extend as far as saying which concepts are instantiated there. Since the main way that we specify which world is fictional is by specifying which concepts are instantiated at it, authorial authority will usually let authors get away with any kind of conceptual claim. But once we have locked onto the world being discussed, the author has no special authority to say which concepts, especially which higher-level concepts like RIGHT or FUNNY or CHAIR are instantiated there.

[12] Again, it is worth noting that I am not ruling out any violation of *Virtue*, just easy violations of it. The point being made in the text is that even a blanket ban on violations would not be a serious restriction on authorial freedom.

(Does it matter much that the distinction between empirical disagreements and conceptual disagreements with which I started might turn out not to rest on very much? Not really. I am trying to explain why we have the attitudes towards fiction that we do, which in turn determines what is true in fiction generally. All that matters is that people generally think that there is something like a conceptual truth/empirical truth distinction, and I think enough people would agree that A and B's disagreement is different in kind from B and C's disagreement to show that is true. If folks are generally wrong about this, if there is no difference in kind between conceptual truths and empirical truths, then our communal theory of truth in fiction will rest on some fairly untenable supports. But it will still be our theory, although any coherent telling of it will have to be in terms of things that are taken to be conceptual truths and things that are taken to be empirical truths.)

This explanation of why authorial authority collapses just when it does yields one fairly startling, and I think true, prediction. I argued above that authors could not easily generate violations of *Virtue*. That this is impossible is compatible with any number of hypotheses about how readers will resolve those impossibilities that authors attempt to slip in. The story here, that authors get to say which world is at issue but not which concepts apply to it, yields the prediction that readers will resolve the tension in favour of the lower-level claims. When given a physical description of a world and an incompatible moral description, we will take the physical description to fix which world is at issue and reduce the moral description to a series of questionable claims about the world. Compare what happens with A, B and C. We take A and B to agree about the world and disagree about concepts, rather than say taking B and C to agree about what the world is like (there's a chair at the heart of the furniture show) and say that A and B disagree about the application of some recognitional concepts. This prediction is borne out in every case discussed in §2. We do not conclude that Craig did not really shoot Jack and Jill, because after all the world at issue is stipulated to be one where he did the right thing. Even more surprisingly, we do not conclude that Quixote's furniture does not look like kitchen utensils, because it consists of a television and an armchair. This is surprising because in *Victory* I never said that the furniture looked like kitchen utensils. The *tacit* low-level claim about appearances is given precedence over the *explicit* high-level claims about which objects populate Quixote's apartment. The theory sketched here predicts that, and supports the solution to the alethic puzzle sketched in §5, which is good news for both the theory and the solution.

It's been a running theme here that the puzzles do not have anything particularly to do with normativity. But some normative concepts raise the kind of issues about authority mentioned here in a particularly striking way. There is always some division of cognitive labour in fiction. The author's role is, among other things, to say which world is being made fictional. The audience's role is, among other things, to determine the artistic merit of the fictional work. On other points there may be some sharing of roles, but this division is fairly absolute. The division threatens to collapse when authors start commenting on the aesthetic quality of words produced by their characters. At the end of *Ivy Day in the Committee Room* Joyce has one character describe a poem just recited by another character as "A fine piece of writing" (Joyce, 1914/2000, 105). Most critics seem to be happy to *accept* the line, because Joyce's poem here really is, apparently, a fine piece of writing. But to me it seems rather jarring, even if it happens to be true. It's easy to feel a similar reaction when characters in a drama praise the words of

another character.[13] This is a special, and especially vivid, illustration of the point I've been pushing towards here. The author gets to describe the world at whichever level of detail she chooses. But once it has been described, the reader has just as much say in which higher-level concepts apply to parts of that world. When the concepts are evaluative concepts that directly reflect on the author, the reader's role rises from being an equal to having more say than the author, just as we normally have less say than others about which evaluative concepts apply to us.

This idea is obviously similar to Yablo's point that we get to decide when grokking concepts apply, not the author. But it isn't quite the same. I think that if any concepts are grokking, most concepts are, so it can't be the case that authors never get to say when grokking concepts apply in their stories. Most of the time authors will get to say which grokking concepts apply, because they have to use them to tell us about the world. What's special about the kind of concepts that cause puzzles is that we get to decide when they apply full stop, but that we get to decide how they apply *given how more fundamental concepts apply*. So the conciliatory version of the relation between my picture here and Yablo's is that I've been filling in, in rather laborious detail, his missing antecedent.

9 Two Hard Cases

The first hard case is suggested by Kendall Walton (1994). Try to imagine a world where the over-riding moral duty is to maximise the amount of nutmeg in the world. If you are like me, you will find this something of a challenge. Now consider a story *Nutmeg* that reads (in its entirety!): "Nobody ever discovered this, but it turned out all along their over-riding moral duty was to maximise the amount of nutmeg in the world." What is true in *Nutmeg*? It seems that there are no violations of *Virtue* here, but it is hard to imagine what is being described.

The second hard case is suggested by Tamar Szabó Gendler (2000). (I'm simplifying this case a little, but it's still hard.) In her *Tower of Goldbach*, God decrees that 12 shall no longer be the sum of two primes, and from this it follows (even in the story) that it is not the sum of 7 and 5. (It is not clear why He didn't just make 5 no longer prime - say the product of 68 and 57. That may have been simpler.) Interestingly, this has practical consequences. When a group of seven mathematicians from one city attempts to join a group of five from another city, they no longer form a group of twelve. Again, two questions. Can we imagine a Goldbachian situation, where 7 and 5 equal not 12? Is it true in Gendler's story that 7 and 5 equal not 12? If we cannot imagine Goldbach's tower, where is the violation of *Virtue*?

First a quick statement of my responses to the two cases then I'll end with my detailed responses. To respond properly we need to tease apart the alethic and imaginative puzzles. I claim that the alethic puzzle only arises when there's a violation of *Virtue*. There's no violation in either story, so there is no alethic puzzle. I think there are independent arguments for this conclusion in both cases. We can't imagine either (if we can't) because any way of filling in the more basic facts leads to violations.

[13] For a while this would happen frequently on the TV series *The West Wing*. President Bartlett would deliver a speech, and afterwards his staffers would congratulate themselves on what a good speech it was. The style of the congratulations was clearly intended to convey the author's belief that the speech they themselves had written was a good speech, not just the characters' beliefs to this effect. When in fact it was a very bad speech, this became very jarring. In later series they would often not show the speeches in question and hence avoid this problem.

It follows from my solution to the alethic puzzle that Nutmegism (Tyler Doggett's name for the principle that we must maximise quantities of nutmeg) could be true in a story. There is no violation in *Nutmeg*, since there are no lower level claims made. Still, the story is very hard to imagine. The reason for this is quite simple. As noted, we cannot just imagine a chair, we have to imagine something more detailed that is a chair in virtue of its more basic properties. (There is no particular more basic property we need imagine, as is shown by the fact that we can imagine a chair just by imagining something with a certain look, or we can imagine a chair in the dark with no visual characteristics. But there is always *something* more basic.) Similarly to imagine a duty, we have to imagine something more detailed, in this case presumably a society or an ecology, in virtue of which the duty exists. But no such possible, or even impossible, society readily springs to mind. So we cannot imagine Nutmegism is true.

But it is hard to see how, or why, this inability should be raised into a restriction on what can be true in a story. One might think that what is wrong with *Nutmeg* is that the fictional world is picked out using high-level predicates. If we extend the story *any way at all*, the thought might go, we will generate a violation of *Virtue*. And that is enough to say that Nutmegism is not true in the story. But actually this isn't quite right. If we extend the story by adding more *moral* claims, there is no duty to minimise suffering, there is no duty to help the poor etc, there are still no violations in the story. The restriction we would have to impose is that there is no way of extending the story to fill out the facts in virtue of which the described facts obtain, without generating a violation. But that looks like too strong a constraint, mostly because if we applied it here, to rule out Nutmegism being true in *Nutmeg*, we would have to apply it to every story written in a higher level language than that of microphysics. It doesn't *seem* true that we have to be able to continue a story all the way to the microphysical before we can be confident that what the author says about, for instance, where the furniture in the room is. So there's no reason to not take the author's word in *Nutmeg*, and since the default is always that what the author says is true, Nutmegism is true in the story.

The mathematical case is more difficult. The argument that 7 and 5 could fail to equal 12 in the story turns on an example by Gregory Currie (1990). (The main conclusions of this example are also endorsed by Byrne (1993).) Currie imagines a story in which the hero refutes Gödel's Incompleteness Theorem. Currie argues that the story could be written in such a way that it is true in the story not merely that everyone believes our hero refuted Gödel, but that she really did. But if it could be true in a story that Gödel's Incompleteness Theorem could be false, then it's hard to see just why it could not be true in a story that a simpler arithmetic claim, say that 7 and 5 make 12, could also be false. Anything that can't be true in a story can't be true in virtue of some feature it has. The only difference between Gödel's Incompleteness Theorem and a simple arithmetic statement appears to be the simplicity of the simple statement. And it doesn't seem possible, or advisable, to work that kind of feature into a theory of truth in fiction.

The core problem here is that how simple a mathematical impossibility is very much a function of the reader's mathematical knowledge and acumen. Some readers probably find the unique prime factorisation theorem so simple and evident that for them a story in which it is false is as crashingly bad as a story in which 7 and 5 do not make 12. For other readers, it is so complex that a story in which it has a counterexample is no more implausible than a story in which Gödel is refuted. I think it cannot be true for the second reader that the unique prime factorisation theorem fails in the story and false for the first reader. That amounts to a kind of relativism about truth in fiction that seems preposterous. But I agree with Currie that

some mathematical impossibilities can be true in a fiction. So I conclude that, whether it is imaginable or not, it could be true in a story that 7 and 5 not equal 12.

I think, however, that it is impossible to imagine that 7 plus 5 doesn't equal 12. Can we explain that unimaginability in the same way we explained why *Nutmeg* couldn't be imagined? I think we can. It seems that the sum of 7 and 5 is what it is in virtue of the relations between 7, 5 and other numbers. It is not primitive that various sums take the values they take. That would be inconsistent with, for example, it being constitutive of addition that it's associative, and associativity does seem to be constitutive of addition. We cannot think about 7, 5, 12 and addition without thinking about those more primitive relations. So we cannot imagine 7 and 5 equally anything else. Or so I think. There's some rather sophisticated, or at least complicated, philosophy of mathematics in the story here, and not everyone will accept all of it. So we should predict that not everyone will think that these arithmetic claims are unimaginable. And, pleasingly, not everyone does. Gendler, for instance, takes it as a data point that *Tower of Goldbach* is imaginable. So far so good. Unfortunately, if the story is true we should also expect that whether people find the story imaginable links up with the various philosophies of mathematics they believe. And the evidence for that is thin. So there may be more work to do here. But there is clearly *a* story that we can tell that handles the case.

David Lewis

David Lewis (1941–2001) was one of the most important philosophers of the 20th Century. He made significant contributions to philosophy of language, philosophy of mathematics, philosophy of science, decision theory, epistemology, meta-ethics and aesthetics. In most of these fields he is essential reading; in many of them he is among the most important figures of recent decades. And this list leaves out his two most significant contributions.

In philosophy of mind, Lewis developed and defended at length a new version of materialism (see the entry on physicalism). He started by showing how the motivations driving the identity theory of mind and functionalism could be reconciled in his theory of mind. He called this an identity theory, though his theory motivated the position now known as analytic functionalism. And he developed detailed accounts of mental content (building on Davidson's interpretationism) and phenomenal knowledge (building on Nemirow's ability hypothesis) that are consistent with his materialism. The synthesis Lewis ended up with is one of the central positions in contemporary debates in philosophy of mind.

But his largest contributions were in metaphysics. One branch of his metaphysics was his Hume-inspired reductionism about the nomological. He developed a position he called "Humean supervenience", the theory that said that there was nothing to reality except the spatio-temporal distribution of local natural properties. And he did this by showing in detail how laws, chances, counterfactual dependence, causation, dispositions and colours could be located within this Humean mosaic. The other branch of his metaphysics was his modal realism. Lewis held that the best theory of modality posited concrete possible worlds. A proposition is possible iff it is true at one of these worlds. Lewis defended this view in his most significant book, *On the Plurality of Worlds*. Alongside this, Lewis developed a new account of how to think about modal properties of individuals, namely counterpart theory, and showed how this theory resolved several long-standing puzzles about modal properties.

1 Lewis's Life and Influence

As we've already seen, part of Lewis's significance came from the breadth of subject matter on which he made major contributions. It is hard to think of a philosopher since Hume who has contributed so much to so many fields. And in all of these cases, Lewis's contributions involved

† Penultimate draft only. Please cite published version if possible. Final version published in Stanford Encyclopaedia of Philosophy. I've learned a lot over the years from talking about Lewis's philosophy with Wolfgang Schwarz. I trust his book (2009) is excellent on all these topics, but unfortunately it's only out in German so far, which I don't read. But a lot of important points are collected on his blog, which is listed under other internet resources. The best book in English on Lewis is Daniel Nolan's *David Lewis* (2005). Without that book, section 7.5 of this entry wouldn't exist, section 6.3 would be unintelligible, and every section would be worse. Much of the biographical information in the introduction is taken from Hájek (2010). Many people helpfully spotted typos and infelicities of expression in earlier versions of this entry. Thanks especially to Zachary Miller for many suggested improvements and revisions. The bibliography is based in large part on a bibliography provided to me by Stephanie Lewis.

defending, or in many cases articulating, a big picture theory of the subject matter, as well as an account of how the details worked. Because of all his work on the details of various subjects, his writings were a font of ideas even for those who didn't agree with the bigger picture. And he was almost invariably clear about which details were relevant only to his particular big picture, and which were relevant to anyone who worked on the subject.

Lewis was born in Oberlin, Ohio in 1941, to two academics. He was an undergraduate at Swarthmore College. During his undergraduate years, his interest in philosophy was stimulated by a year abroad in Oxford, where he heard J. L. Austin's final series of lectures, and was tutored by Iris Murdoch. He returned to Swarthmore as a philosophy major, and never looked back. He studied at Harvard for his Ph.D., writing a dissertation under the supervision of W. V. O. Quine that became his first book, *Convention*. In 1966 he was hired at UCLA, where he worked until 1970, when he moved to Princeton. He remained at Princeton until his death in 2001. While at Harvard he met his wife Stephanie. They remained married throughout Lewis's life, jointly attended numerous conferences, and co-authored three papers. Lewis visited Australia in 1971, 1975, every year from 1979 to 1999, and again shortly before his death in 2001.

Lewis was a Fellow of the American Academy of Arts and Sciences, a Corresponding Fellow of the British Academy, and an Honorary Fellow of the Australian Academy of the Humanities. He received honorary doctorates from the University of Melbourne, the University of York in England, and Cambridge University. His Erd os number was 3.

Lewis published four books: *Convention* (1969a), *Counterfactuals* (1973b), *On the Plurality of Worlds* (1986b) and *Parts of Classes* (1991). His numerous papers have been largely collected in five volumes: *Philosophical Papers Vol. I* (1983c), *Philosophical Papers Vol. II* (1986c), *Papers in Philosophical Logic* (1998), *Papers in Metaphysics and Epistemology* (1999a) and *Papers in Social Philosophy* (2000). This entry starts with a discussion of Lewis's first two books, then looks at his contributions to philosophy of mind. Sections 5 and 6 are on his metaphysics, looking in turn at Humean Supervenience and modal realism. Section 7 looks very briefly at some of the many works that aren't been covered in the previous five categories.

2 Convention

David Lewis's first book was *Convention* (1969a, note that all citations are to works by David Lewis, unless explicitly stated otherwise). It was based on his Harvard Ph. D. thesis, and published in 1969. The book was an extended response to the arguments of Quine and others than language could not be conventional. Quine's argument was that conventions are agreements, and agreements require language, so language must be prior to any convention, not a consequence of a convention. Lewis's response is to deny that conventions require anything like an agreement. Rather, on his view, conventions are regularities in action that solve co-ordination problems. We can stumble into such a regularity without ever agreeing to do so. And such a regularity can persist simply because it is in everyone's best interest that it persist.

2.1 Analysis of Convention

Lewis viewed conventions as solutions to co-ordination problems (see Section 3.2 of the entry on convention). His thinking about these problems was heavily influenced by Thomas Schelling's work on co-operative games in *The Strategy of Conflict* (Schelling, 1960). Many of the key ideas in Lewis's book come from game theory.

The simplest cases in which conventions arise are ones where we are repeatedly playing a game that is purely co-operative, i.e. the payoffs to each agent are the same, and there are multiple equilibria. In such a case, we may well hope for the equilibrium to persist. At the very least, we will prefer the persistence of the equilibrium to any one person deviating from it. And we will have this preference even if we would prefer, all things considered, to be in some other equilibrium state. In such a case, there may well be a practice of continuing to play one's part in the equilibrium that has been reached. This is a regularity in action—it involves making moves in the repeated game. Given that everyone else is following the regularity, each agent has a reason to follow the regularity; otherwise it wouldn't be an equilibrium. But if other agents acted differently, agents would not be interested in following the regularity, since there are alternative equilibria. Because these three conditions are met, Lewis argued that the practice is really a convention, even if there was never any explicit agreement to continue it.

The case we started with was restricted in two important ways. First, the case involved games that were perfectly repeated. Second, it involved games where the payoffs were perfectly symmetric. Lewis's theory of convention involved getting rid of both restrictions.

Instead of focussing on repeated co-ordination problems, Lewis just focussed on repeated situations which collectively constitute a co-ordination problem. Lewis does not identify situations with games. A repeated situation may come in different 'versions', each of which is represented by a different game. For example, it may be that the costs of performing some kind of action differ on different occasions, so the formal game will be different, but the differences are small enough that it makes sense to have a common practice. And Lewis does not require that there be identity of interests. In *Convention* he does require that there be large overlap of interests, but this requirement does not do much work, and is abandoned in later writing. With those requirements weakened, we get the following definition of convention.

> A regularity R in the behaviour of members of a population P when they are agents in a recurrent situation S is a *convention* if and only if it is true that, and it is common knowledge in P that, in almost any instance of S among members of P,
>
> 1. almost everyone conforms to R;
> 2. almost everyone expects everyone else to conform to R;
> 3. almost everyone has approximately the same preferences regarding all possible combinations of actions;
> 4. almost everyone prefers that any one more conform to R, on condition that almost everyone conform to R;
> 5. almost everyone would prefer that any one more conform to R', on condition that almost everyone conform to R',
>
> where R' is some possible regularity in the behaviour of members of P in S, such that almost no one in almost any instance of S among members of P could conform to both R' and to R. (Lewis, 1969a, 78)

This is clearly a vague definition, with many 'almost's scattered throughout. But Lewis, characteristically, thought this was a feature not a bug of the view. Our intuitive notion of a

convention is vague, and any analysis of it should capture the vagueness. The idea that analyses of imprecise folk concepts should be imprecise recurs throughout Lewis's career.

The notion of 'common knowledge' that Lewis is working with here is not the standard modern notion. Lewis does not require that everyone know that everyone know etc., that all of these conditions hold. Rather, when Lewis says that it is common knowledge that p, he means that everyone has a reason to believe that p, and everyone has a reason to believe everyone has a reason to believe that p, and everyone has a reason to believe that everyone has a reason to believe everyone has a reason to believe that p, and so on. That people act on these reasons, or are known to act on these reasons, to form beliefs is unnecessary. And that the beliefs people would get if they acted on their reasons are true is also not part of the view. Hence it is necessary to specify truth as well as common belief in the definition.

Lewis argues that this definition captures many of our ordinary conventions, such as the convention of driving on the right side of the road in the United States, the convention of taking certain pieces of paper as payments for debts, and, most importantly, the conventions governing the use of language.

2.2 Conventions of Language

In the final chapter of *Convention*, Lewis gives his theory of what it is for a community to speak a language (see the section on conventional theories of meaning in the entry on convention), i.e., for a community to have adopted one language as their language by convention. Lewis individuates languages largely by the truth conditions they assign to sentences. And his account of truth conditions is given in terms of possible worlds. So the truth condition of an indicative sentence is the set of possible worlds in which it is true. Somewhat more abnormally, Lewis takes the truth condition for an imperative to be the set of possible worlds in which the imperative is obeyed. (The account of language in *Convention* covers many different moods, but we will focus here on the account of indicatives.)

The focus on truth conditions is not because Lewis thinks truth conditions are all that there are to languages. He acknowledges that languages also have 'grammars'. A grammar, in Lewis's sense, is a lexicon (i.e. a set of elementary constituents, along with their interpretation), a generative component (i.e. rules for combining constituents into larger constituents), and a representing component (i.e. rules for verbally expressing constituents). Lewis's preferred interpretations are functions from possible worlds to extensions. So we can sensibly talk about the meaning of a non-sentential constituent of the language, but these meanings are derived from the truth conditions of sentences, rather than determining the meanings of sentences. That's because, as we'll see, what the conventions of language establish in the first instance are truth conditions for entire messages, i.e., sentences.

Given this understanding of what a language is, Lewis goes on to say what it is for a population to speak a language. One natural approach would be to say that speakers and hearers face a co-ordination problem, and settling on one language to communicate in would be a solution to that problem. When Lewis is analysing signalling, that is the approach he takes. But he doesn't think it will work for language in general. The reason is that he takes conventions to be regularities in action, and it is hard to say in general what actions are taken by hearers.

So instead Lewis says that a population P speaks a language L iff there is a convention of speaking truthfully in L that persists amongst P. The parties to the co-ordination problem (and

the convention that solves it) are the different people who want to communicate in P. They solve their problem by speaking truthfully (on the whole) in L.

It might be wondered whether it could really be a convention to speak truthfully in L. After all, there is no obvious alternative to speaking truthfully. As Lewis points out, however, there are many natural alternatives to speaking truthfully in L; we could speak truthfully in L' instead. The existence of alternative languages makes our use of L conventional. And the convention can be established, and persist, without anyone agreeing to it.

2.3 Later Revisions

In "Languages and Language" (1975b), Lewis makes two major revisions to the picture presented in *Convention*. He changes the account of what a convention is, and he changes the account of just what convention must obtain in order for a population to speak a language.

There are two changes to the account of convention. First, Lewis now says that conventions may be regularities in action and belief, rather than just in action. Second, he weakens the third condition, which was approximate sameness of preferences, to the condition that (almost) each agent has a reason to conform when they believe others conform. The reason in question may be a practical reason, when conformity requires action, or an epistemic reason, when convention requires belief.

In *Convention*, the conventions that sustained language were regularities amongst speakers. As we noted, it would be more natural to say that the conventions solved co-ordination problems between speakers of a language and their hearers. That is what the new account of what it is for a population to speak a language does. The population P speaks the language L iff there are conventions of truthfulness and trust in L. Speakers are truthful in L iff they only utter sentences they believe are true sentences of L. Hearers are trusting in L iff they take the sentences they hear to be (generally) true sentences of L.

The old account took linguistic conventions to be grounded in co-ordination between speakers generally. We each communicate in English because we think we'll be understood that way given everyone else communicates that way, and we want to be understood. In the new account there is still this kind of many-way co-ordination between all the speakers of a language, but the most basic kind of co-ordination is a two-way co-ordination between individual speakers, who want to be understood, and hearers, who want to understand. This seems like a more natural starting point. The new account also makes it possible for someone to be part of a population that uses a language even if they don't say anything because they don't have anything to say. As long as they are trusting in L, they are part of the population that conforms to the linguistic regularity.

John Hawthorne (1990) argued that Lewis's account cannot explain the intuitive meaning of very long sentences. While not accepting all of Hawthorne's reasons as to why very long sentences are a problem, in "Meaning Without Use: Reply to Hawthorne" (1992) Lewis agreed that such sentences pose a problem to his view. To see the problem, let L be the function from each sentence of English to its intuitive truth condition, and let L^* be the restriction of that function to sentences that aren't very long. Arguably we do not trust speakers who utter very long sentences to have uttered truths, under the ordinary English interpretation of their sentences. We think, as Lewis said, that such speakers are "trying to win a bet or set a record, or feigning madness or raving for real, or doing it to annoy, or filibustering, or making an experiment to test the limits of what it is humanly possible to say and mean." (1992, 108)

That means that while there may be a convention of truthfulness and trust in L^*, there is no convention of trust in L in its full generality. So the "Languages and Language" theory implies that we speak L^*, not L, which is wrong.

Lewis's solution to this puzzle relies on his theory of natural properties, described below in Section 4.6. He argues that some grammars (in the above sense of grammar) are more natural than others. By default, we speak a language with a natural grammar. Since L has a natural grammar, and L^* doesn't, other things being equal, we should be interpreted as speaking L rather than L^*. Even if other things are not quite equal, i.e. we don't naturally trust speakers of very long sentences, if there is a convention of truthfulness and trust in L in the vast majority of verbal interactions, and there is no other language with a natural grammar in which there is a convention of truthfulness and truth, then the theory will hold, correctly, that we do speak L.

3 Counterfactuals

David Lewis's second book was *Counterfactuals* (1973b). Counterfactual conditionals were important to Lewis for several reasons. Most obviously, they are a distinctive part of natural language and it is philosophically interesting to figure out how they work. But counterfactuals would play a large role in Lewis's metaphysics. Many of Lewis's attempted reductions of nomic or mental concepts would be either directly in terms of counterfactuals, or in terms of concepts (such as causation) that he in turn defined in terms of counterfactuals. And the analysis of counterfactuals, which uses possible worlds, would in turn provide motivation for believing in possible worlds. We will look at these two metaphysical motivations in more detail in section 4, where we discuss the relationship between counterfactuals and laws, causation and other high-level concepts, and in section 5, where we discuss the motivations for Lewis's modal metaphysics.

3.1 Background

To the extent that there was a mid-century orthodoxy about counterfactual conditionals, it was given by the proposal in Nelson Goodman (1955). Goodman proposed that counterfactual conditionals were a particular variety of strict conditional. To a first approximation, If it were the case that p, it would be the case that q (hereafter $p \mathrel{\Box\!\!\rightarrow} q$) is true just in case Necessarily, either p is false or q is true, i.e. $\Box(p \supset q)$. Goodman realised that this wouldn't work if the modal 'necessarily' was interpreted unrestrictedly. He first suggested that we needed to restrict attention to those possibilities where all facts 'co-tenable' with p were true. More formally, if S is the conjunction of all the co-tenable facts, then $p \mathrel{\Box\!\!\rightarrow} q$ is true iff $\Box((p \wedge S) \supset q)$.

Lewis argued that this could not be the correct set of truth conditions for $p \mathrel{\Box\!\!\rightarrow} q$ in general. His argument was that strict conditionals were in a certain sense indefeasible. If a strict conditional is true, then adding more conjuncts to the antecedent cannot make it false. But intuitively, adding conjuncts to the antecedent of a counterfactual can change it from being true to false. Indeed, intuitively we can have long sequences of counterfactuals of ever increasing strength in the antecedent, but with the same consequent, that alternate in truth value. So we can imagine that (3.1) and (3.3) are true, while (3.2) and (3.4) are false.

(3.1) If Smith gets the most votes, he will be the next mayor.

(3.2) If Smith gets the most votes but is disqualified due to electoral fraud, he will be the next mayor.
(3.3) If Smith gets the most votes, but is disqualified due to electoral fraud, then launches a military coup that overtakes the city government, he will be the next mayor.
(3.4) If Smith gets the most votes, but is disqualified due to electoral fraud, then launches a military coup that overtakes the city government, but dies during the coup, he will be the next mayor.

If we are to regard $p \mathrel{\Box\!\!\rightarrow} q$ as true iff $\Box((p \wedge \mathrm{S}) \supset q)$, then the S must vary for different values of p. More seriously, we have to say something about *how* S varies with variation in p. Goodman's own attempts to resolve this problem had generally been regarded as unsuccessful, for reasons discussed in Bennett (1984). So a new solution was needed.

3.2 Analysis

The basic idea behind the alternative analysis was similar to that proposed by Robert Stalnaker (1968). Let's say that an A-world is simply a possible world where A is true. Stalnaker had proposed that $p \mathrel{\Box\!\!\rightarrow} q$ was true just in case the most similar p-world to the actual world is also a q-world. Lewis offered a nice graphic way of thinking about this. He proposed that we think of similarity between worlds as a kind of metric, with the worlds arranged in some large-dimensional space, and more similar worlds being closer to each other than more dissimilar worlds. Then Stalnaker's idea is that the closest p-world has to be a q-world for $p \mathrel{\Box\!\!\rightarrow} q$ to be true. Lewis considered several ways of filling out the details of this proposal, three of which will be significant here.

First, he rejected Stalnaker's presupposition that there is a most similar p-world to actuality. He thought there might be many worlds which are equally similar to actuality, with no p-world being more similar. Using the metric analogy suggested above, these worlds all fall on a common 'sphere' of worlds, where the centre of this sphere is the actual world. In such a case, Lewis held that $p \mathrel{\Box\!\!\rightarrow} q$ is true iff all the p-worlds on this sphere are q-worlds. One immediate consequence of this is that Conditional Excluded Middle, i.e., $(p \mathrel{\Box\!\!\rightarrow} q) \vee (p \mathrel{\Box\!\!\rightarrow} \neg q)$ is not a theorem of counterfactual logic for Lewis, as it was for Stalnaker.

Second, he rejected the idea that there must even be a sphere of closest p-worlds. There might, he thought, be closer and closer p-worlds without limit. He called the assumption that there was a sphere of closest worlds the "Limit Assumption", and noted that we could do without it. The new truth conditions are that $p \mathrel{\Box\!\!\rightarrow} q$ is true at w iff there is a $p \wedge q$-world closer to w than any $p \wedge \neg q$-world.

Third, he considered dropping the assumption that w is closer to itself than any other world, or even the assumption that w is among the worlds that are closest to it. When we think in terms of similarity (or indeed of metrics) these assumptions seem perfectly natural, but some philosophers have held that they have bad proof theoretic consequences. Given the truth conditions Lewis adopts, the assumption that w is closer to itself than any other world is equivalent to the claim that $p \wedge q$ entails $p \mathrel{\Box\!\!\rightarrow} q$, and the assumption that w is among the worlds that are closest to it is equivalent to the claim that $p \mathrel{\Box\!\!\rightarrow} q$ and p entail q. The first of these entailments in particular has been thought to be implausible. But Lewis ultimately decided to endorse it, in large part because of the semantic model he was using. When we

don't think about entailments, and instead simply ask ourselves whether any other world could be as similar to w as w is to itself, the answer seems clearly to be no.

As well as offering these semantic models for counterfactuals, in the book Lewis offers an axiomatisation of the counterfactual logic he prefers (see the section on the Logic of Ontic Conditionals in the entry the logic of conditionals), as well as axiomatisations for several other logics that make different choices about some of the assumptions we've discussed here. And he has proofs that these axiomatisations are sound and complete with respect to the described semantics.

He also notes that his preferred counterfactual logic invalidates several familiar implications involving conditionals. We already mentioned that strengthening the antecedent, the implication of $(p \wedge r) \boxrightarrow q$ by $p \boxrightarrow q$, is invalid on Lewis's theory, and gave some natural language examples that suggest that it should be invalid. Lewis also shows that contraposition, the implication of $q \boxrightarrow p$ by $p \boxrightarrow q$, and conditional syllogism, the implication of $p \boxrightarrow r$ by $p \boxrightarrow q$ and $q \boxrightarrow r$, are invalid on his model, and gives arguments that they should be considered invalid.

3.3 Similarity

In *Counterfactuals*, Lewis does not say a lot about similarity of worlds. He has some short arguments that we can make sense of the notion of two worlds being similar. And he notes that on different occasions we may wish to use different notions of similarity, suggesting a kind of context dependency of counterfactuals. But the notion is not spelled out in much more detail.

Some reactions to the book showed that Lewis needed to say more here. Kit Fine (1975a) argued that given what Lewis had said to date, (3.5) would be false, when it should be true.

(3.5) If Richard Nixon had pushed the button, there would have been a nuclear war.

('The button' in question is the button designed to launch nuclear missiles.) The reason it would be false is that a world in which the mechanisms of nuclear warfare spontaneously failed but then life went on as usual, would be more similar, all things considered, to actuality than a world in which the future consisted entirely of a post-nuclear apocalypse.

In "Counterfactual Dependence and Time's Arrow" (1979b), Lewis responded by saying more about the notion of similarity. In particular, he offered an algorithm for determining similarity in standard contexts. He still held that the particular measure of similarity in use on an occasion is context-sensitive, so there is no one true measure of similarity. Nevertheless there is, he thought, a default measure that we use unless there is a reason to avoid it. Here is how Lewis expressed this default measure.

1. It is of the first importance to avoid big, widespread, diverse violations of law.
2. It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
3. It is of the third importance to avoid even small, localized, simple violations of law.
4. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly. (1979b, 47-48)

Lewis argues that by this measure, worlds in which the mechanisms of nuclear warfare spontaneously fail will be less similar to the actual world than the post-nuclear apocalypse. That's because the failure of those mechanisms will either lead to divergence from the actual world (if they fail partially) or widespread, diverse violations of law (if they fail completely). In the former case, there's a violation of law that isn't made up for in an increase in how much spatio-temporal match we get. In the latter case the gain we get in similarity is only an expansion of the spatio-temporal region throughout which perfect match of particular fact prevails, but that doesn't help in getting us closer to actuality if we've added a big miracle. So in fact the nearest worlds are ones where a nuclear war occurs, and (3.5) is true.

One way to see the effects of Lewis's ordering is to work through its implication for an important class of cases. When the antecedent of a counterfactual is about the occurrence or non-occurrence of a particular event E at time t, the effect of these rules is to say that the nearest worlds are the worlds where the following claims all hold, with t^* being as late as possible.

- There is an exact match of particular fact with actuality up to t^*.
- There is a small, localized law violation at t^*.
- There is exact conformity to the laws of actuality after t^*.
- The antecedent is true.

So we find a point just before t where we can make the antecedent true by making a small law violation, and let the laws take over from there. There is something intuitively plausible about this way of viewing counterfactuals; often we do aim to talk about what would have happened if things had gone on in accordance with the laws, given a starting point slightly different from the one that actually obtained.

Jonathan Bennett (2003) notes that when the antecedent of a conditional is not about a particular event, Lewis's conditions provide the wrong results. For instance, if the antecedent is of the form *If one of these events had not happened*, then Lewis's rules say that the nearest world where the antecedent is true is always the world where the most recent such event did not happen. But this does not seem to provide intuitively correct truth conditions for such conditionals. This need not bother Lewis's larger project. For one thing, Lewis was not committed to there being a uniform similarity metric for all counterfactuals. Lewis could say that his default metric was only meant to apply to cases where the antecedent was about the happening or non-happening of a particular event at a particular time, and it wouldn't have seriously undermined his larger project. Indeed, as we'll see in Section 5.2 below, the counterfactuals he was most interested in, and for which these criteria of similarity were devised, did have antecedents concerning specific events.

4 Philosophy of Mind

In "Reduction of Mind" (1994b), David Lewis separates his contributions to philosophy of mind into two broad categories. The first category is his reductionist metaphysics. From his first published philosophy paper, "An Argument for the Identity Theory" (1966), Lewis defended a version of the mind-brain identity theory (see the entry on the identity theory of mind). As he makes clear in "Reduction of Mind", this became an important part of his global reductionism. We'll look at his metaphysics of mind in sections 4.1–4.3.

The second category is his interpretationist theory of mental content. Following Donald Davidson in broad outlines, Lewis held that the contents of a person's mental states are those contents that a radical interpreter would interpret them as having, assuming the interpreter went about their task in the right way. Lewis had some disagreements with Davidson (and others) over the details of interpretationism, but we won't focus on those here. What we will look at are two contributions that are of interest well beyond interpretationism, indeed beyond theories of mental content. Lewis held that mental contents are typically *properties*, not *propositions*. And he held that a theory of mental content requires an *inegalitarian* theory of properties. We'll look at his theory of content in sections 4.4–4.6.

4.1 Ramsey Sentences

The logical positivists faced a hard dilemma when trying to make sense of science. On the one hand, they thought that all meaningful talk was ultimately talk about observables. On the other hand, they respected science enough to deny that talk of unobservables was meaningless. The solution was to 'locate' the unobservables in the observation language; in other words, to find a way to reduce talk of unobservables to talk about observables.

Lewis didn't think much of the broader positivist project, but he was happy to take over some of their technical advances in solving this location problem. Lewis noted that this formal project, the project of trying to define theoretical terms in an already understood language, was independent of the particular use we make of it. All that really matters is that we have some terms introduced by a new theory, and that the new theory is introduced in a language that is generally understood. In any such case it is an interesting question whether we can extract the denotation of an introduced term from the theory used to introduce it.

The term-introducing theory could be a scientific theory, such as the theory that introduces terms like 'electron', and the language of the theory could be observation language. Or, more interestingly, the term-introducing theory could be folk psychology, and the language of the theory could be the language of physics. If we have a tool for deriving the denotations of terms introduced by a theory, and we have a way of treating folk psychology as a theory (i.e., a conjunction of sentences to which folk wisdom is committed), we can derive the denotations of terms like 'belief', 'pain', and so on using this theory. Some of Lewis's important early work on the metaphysics of mind was concerned with systematising the progress positivists, especially Ramsey and Carnap, had made on just this problem. The procedure is introduced in "An Argument for the Identity Theory", "Psychophysical and Theoretical Identifications" (1972) and "How to Define Theoretical Terms" (1970b). There are important later discussions of it in "Reduction of Mind" and "Naming the Colours" (1997c), among many others.

In the simplest case, where we have a theory T that introduces one new name t, Lewis says that t denotes the x such that T[x], where T[x] is the sentence we get by (a) converting T to a single sentence, perhaps a single long conjunction, and (b) replacing all occurrences of t with the variable x. That is, if there is a unique x such that T[x], t denotes it, and t is denotationless otherwise. (Note that it isn't *meaningless*, but it is *denotationless*.)

The simplest case is not fully general in a few respects. First, theories often introduce many terms simultaneously, not just one. So the theory might introduce new terms t_1, t_2, ..., t_n. No problem, we can just quantify over n-tuples, where n is the number of new terms introduced. So instead of looking at $\exists_1 x$ T[x], where $\exists_1$ means 'exists a unique' and x is an individual variable, we look at $\exists_1 \mathbf{x}$ T[$\mathbf{x}$], where $\mathbf{x}$ is a variable that ranges over n-tuples, and T[$\mathbf{x}$] is the

sentence you get by replacing t_1 with the first member of **x**, t_2 with the second member of **x**, ..., and t_n with the n^{th} member of **x**. Although this is philosophically very important, for simplicity I'll focus here on the case where a single theoretical term is to be introduced.

The simplest case is not general in another, more important, respect. Not all theoretical terms are names, so it isn't obvious that we can quantify over them. Lewis's response, at least in the early papers, is to say we can always *replace* them with names that amount to the same thing. So if T says that all *F*s are *G*s, and we are interested in the term 'G', then we'll rewrite T so that it now says *G*ness is a property of all *F*s. In the early papers, Lewis says that this is a harmless restatement of T, but this isn't correct. Indeed, in later papers such as "Void and Object" (2004c) and "Tensing the Copula" (2002a) Lewis notes that some predicates don't correspond to properties or relations. There is no property of being non-self-instantiating, for instance, though we can predicate that of many things. In those cases the rewriting will not be possible. But in many cases, we can rewrite T, and then we can quantify into it.

The procedure here is often called Ramsification, or Ramseyfication. (Both spellings have occurred in print. The first is in the title of Braddon-Mitchell and Nola (1997), the second in the title of Melia and Saatsi (2006).) The effect of the procedure is that if we had a theory T which was largely expressed in the language O, except for a few terms t_1, t_2, ..., t_n, then we end up with a theory expressed entirely in the O-language, but which, says Lewis, has much the same content. Moreover, if the converted theory is true, then the T-terms can be defined as the substitutends that make the converted sentence true. This could be used as a way of eliminating theoretical terms from an observation language, if O is the observation language. Or it could be a way of understanding theoretical terms in terms of natural language, if O is the old language we had before the theory was developed.

In cases where there is a unique x such that T[x], Lewis says that t denotes that x. What if there are many such x? Lewis's official view in the early papers is that in such a case t does not have a denotation. In "Reduction of Mind", Lewis retracted this, and said that in such a case t is indeterminate between the many values. In "Naming the Colours" he partially retracts the retraction, and says that t is indeterminate if the different values of x are sufficiently similar, and lacks a denotation otherwise.

A more important complication is the case where there is no realiser of the theory. Here it is important to distinguish two cases. First, there is the case where the theory is very nearly realised. That is, a theory that contains enough of the essential features of the original theory turns out to be true. In that case we still want to say that the theory manages to provide denotations for its new terms. Second, there are cases where the theory is a long way from the truth. The scientific theory of phlogiston, and the folk theory of witchcraft, are examples of this. In this case we want to say that the terms of the theory do not denote.

As it stands, the formal theory does not have the resources to make this distinction. But this is easy to fix. Just replace the theory T with a theory T*, which is a long disjunction of various important conjuncts of T. So if T consisted of three claims, p_1 p_2 and p_3, and it is close enough to true if two of them are true, then T* would be the disjunction $(p_1 \wedge p_2) \vee (p_1 \wedge p_3) \vee (p_2 \wedge p_3)$. Lewis endorses this method in "Psychophysical and Theoretical Identifications" The disjuncts are propositions that are true in states that would count as close enough to the world as described by T that T's terms denote. Note that in a real-world case, some parts of T will be more important than others, so we won't be able to just 'count the conjuncts'. Still, we

should be able to generate a plausible T* from T. And the rule in general is that we apply the above strategy to T* rather than T to determine the denotation of the terms.

4.2 Arguing for the Identity Theory

Lewis's first, and most important, use of Ramsification was to argue for the mind-brain identity theory, in "An Argument for the Identity Theory". Lewis claims in this paper that his argument does not rely on parsimony considerations. The orthodox argument for the identity theory at the time, as in e.g. J. J. C. Smart (1959), turned on parsimony. The identity theory and dualism explain the same data, but the dualist explanation involves more ontology than the identity theory explanation. So the identity theory is preferable. Lewis says that this abductive step is unnecessary. (He even evinces concern that it is unsound.) Lewis offers instead an argument from the causal efficacy of experience. The argument is something like the following. (I've given the argument that pains are physical, a similar argument can be given for any other kind of experience.)

1. Pains are the kind of thing that typically have such-and-such physical causes and such-and-such physical effects, where the 'such-and-such's are filled in by our folk theory of pain.
2. Since the physical is causally closed, the things that have such-and-such physical causes and such-and-such physical effects are themselves physical.
3. So, pains are physical.

The first premise is analytically true; it follows from the way we define theoretical terms. The second premise is something we learn from modern physics. (It isn't clear, by the way, that we can avoid Smart's parsimony argument if we really want to argue for premise 2.) So the conclusion is contingent, since modern physics is contingent, but it is well-grounded. Indeed, if we change the second premise a little, drawing on neurology rather than physics, we can draw a stronger conclusion, one that Lewis draws in "Psychophysical and Theoretical Identifications".

1. Pains are the kind of thing that typically have such-and-such physical causes and such-and-such physical effects, where the 'such-and-such's are filled in by our folk theory of pain.
2. Neural state N is the state that has such-and-such physical causes and such-and-such physical effects.
3. So, pains are instances of neural state N.

So, at least in the second argument, Lewis is defending a kind of identity theory. Pains just are instances of neural states. I'll finish up this survey of Lewis's metaphysics of mind with a look at two complications to this theory.

4.3 Madmen and Martians

Pain is defined by its causal role. Central to that role is that we are averse to pain, and try to avoid it. But not all of us do. Some of us seek out pain. Call them madmen. A good theory of pain should account for the possibility of madmen.

The simplest way to account for madmen would be to simply identify pain with a neural state. So Lewis's identity theory is well-placed to deal with them. But there is a complication. Not every creature in the universe who is in pain has the same neural states as us. It is at least possible that there are creatures in which some silicon state S plays the pain role. That is, the creatures are averse to S, they take S to be an indicator of bodily damage, and so on. Those creatures are in pain whenever they are in state S. Call any such creature a Martian. A simple identification of pain with neural state N will stipulate that there couldn't be any Martians. That would be a bad stipulation to make.

The possibility of madmen pushes us away from a simple functional definition of pain. Some creatures have pains that do not play the pain role. The possibility of Martians pushes us away from a purely neural definition of pains. Some creatures have pains that are not like our neural pain states. Indeed, some of them might have pains without having any neural states at all. Lewis's way of threading this needle is to say that pains, like all mental states, are defined for kinds of creatures. Pains in humans are certain neural states. They are the neural states that (typically, in humans) have the functional role that we associate with pain. In other kinds of creatures pains are other states that (typically, in those creatures) play the pain role. The details of these views are worked out in "Mad Pain and Martian Pain" (1980a).

4.4 Interpretationism

In a recent *Philosophical Review* paper, J. Robert G. Williams describes the theory of content that Lewis endorses as 'interpretationist' (Williams, 2007). It is a good name. It's a platitude that the content of someone's mental states is the interpretation of those states that a good interpreter would make. If it were otherwise, the interpreter wouldn't be good. What's distinctive about interpretationism is the direction of explanatory priority. What makes a person's states have the content they do is that a good interpreter would interpret them that way. This is the core of Lewis's theory of mental content.

Put this broadly, Lewis's position is obviously indebted to Donald Davidson's work, and Lewis frequently acknowledges the debt. But Lewis differs from Davidson in several respects. I'll briefly mention four of them here, then look at two substantial changes in the next two sections. (The primary sources for the discussion in this section are "Radical Interpretation" (1974a) and especially its appendices in *Philosophical Papers: Volume I* (1983c), and "Reduction of Mind".)

First, Lewis does not think that part of being a good interpreter is that we interpret the subject so that as many of their beliefs as possible come out true. Rather, he thinks we should interpret someone so that as many of their beliefs as possible come out rational. If the subject is surrounded by misleading evidence, we should interpret her as having false beliefs rather than lucky guesses.

Second, Lewis does not give a particularly special place to the subject's verbal behaviour in interpreting them. In particular, we don't try to (radically) interpret the subject's language and then use that to interpret their mind. Rather, Lewis follows Grice (among others) in

taking mental content to be metaphysically primary, and linguistic content to be determined by mental states (see the section on meaning in the entry on Grice).

Third, Lewis believes in narrow content. Indeed, there is a sense in which he thinks narrow content is primary. He disagrees with Davidson, and several others, when he holds that Swampman has contentful states. And he thinks that we share many beliefs (most clearly metalinguistic beliefs) with denizens of Twin Earth.

Finally, Lewis's theory of mental content, like his theory of mind in general, is anti-individualistic. What matters is the functional role that a state typically has in creatures of a certain kind, not what role it has in this creature. So there might be a madman who does not attempt to get what they desire. A pure functionalist may say that such a person has no desires, since desires, by definition, are states that agents attempt to satisfy. Lewis says that as long as this state typically leads to satisfaction-attempts in creatures of this kind, it is a desire. Indeed, if it typically leads to attempts to get *X*, it is a desire for *X*, even if little about the role the state plays in this agent would suggest it is a desire for *X*.

4.5 *De Se* Content

Some of our beliefs and desires are about specific individuals. I might, for instance, believe that BW is a crook and desire that he be punished. Some of our beliefs and desires are self-directed. I might, for instance, believe that I am not a crook and desire that I not be punished. If I know that I am BW, then I should not have all of those beliefs and desires. But I might be ignorant of this. In some circumstances (e.g., amnesia, or receiving deceptive information about your identity) it is no sign of irrationality to not know who you are. And if you don't know you are *X*, you may ascribe different properties to yourself and to *X*.

Lewis's way of handling this problem was exceedingly simple. His original version of interpretationism had it that belief-states were ultimately probability distributions over possible worlds, and desire-states were ultimately utility functions, again defined over possible worlds. In "Attitude *De Dicto* and *De Se*" (1979a), he argued that this isn't correct. Beliefs and desires are, at the end of the day, probability and utility functions. (Or at least they are approximations to those functions.) But they are not defined over possible worlds. Rather, they are defined over possible individuals.

What that means for belief and desire is easiest to express using the language of possible worlds. The standard view is that propositions are (or at least determine) sets of possible worlds, and that the content of a belief is a proposition. To believe something then is to locate yourself within a class of possible worlds; to believe that you inhabit one of the worlds at which the proposition is true. Lewis's view is that properties are (or at least determine) sets of possible individuals, and that the content of a belief is a property. To believe something then is to locate yourself within a class of possible individuals; to believe that you are one of the individuals with the property. More simply, beliefs are the self-ascriptions of properties.

Within this framework, it is easy to resolve the puzzles we addressed at the top of the section. If I believe that BW is a crook, I self-ascribe the property of inhabiting a world in which BW is a crook. (On Lewis's theory, beliefs that are not explicitly self-locating will be beliefs about which world one is in.) If I believe I am not a crook, I self-ascribe the property of not being a crook. Since there are possible individuals who are (a) not crooks but (b) in worlds where BW is a crook, this is a consistent self-ascription. Indeed, I may even have strong evidence that I have both of these properties. So there is no threat of inconsistency, or even irrationality here.

Lewis's suggestion about how to think of self-locating mental states has recently been very influential in a variety of areas. Adam Elga (2000a, 2004) has extensively investigated the consequences of Lewis's approach for decision theory. Andy Egan (2007) has developed a novel form of semantic relativism using Lewis's approach as a model. Daniel Nolan (2007) has recently argued that Lewis's approach is less plausible for desire than for belief, and Robert Stalnaker (2008) argues that the view makes the wrong judgments about sameness and difference of belief across agents and times.

4.6 Natural Properties

One classic problem for interpretationism is that our dispositions massively underdetermine contents. I believe that (healthy) grass is green. But for some interpretations of 'grue', ascribing to me the belief that grass is grue will fit my dispositions just as well. As Lewis points out towards the end of "New Work For a Theory of Universals" (1983b), if we are allowed to change the interpretations of my beliefs and desires at the same time, the fit can be made even better. This looks like a problem for interpretationism.

The problem is of course quite familiar. In different guises it is Goodman's grue/green problem, Kripkenstein's plus/quus problem, Quine's gavagai problem, and Putnam's puzzle of the brain in a vat with true beliefs (Goodman, 1955; Wittgenstein, 1953; Kripke, 1982; Quine, 1960; Putnam, 1981). One way or another it has to be solved.

Lewis's solution turns on a metaphysical posit. Some properties, he says, are more *natural* than others. The natural properties are those that, to use an ancient phrase, carve nature at the joints. They make for objective resemblance amongst the objects that have them, and objective dissimilarity between things that have them and those that lack them. The natural properties, but not in general the unnatural properties, are relevant to the causal powers of things. Although science is in the business of discovering which natural properties are instantiated, when Lewis talks about natural properties he doesn't mean properties given a special role by nature. It is not a contingent matter which properties are natural, because it isn't a contingent matter which properties make for objective similarity.

Some properties are perfectly natural. Other properties are less natural, but not all unnatural properties are alike. Green things are a diverse and heterogeneous bunch, but they are more alike than the grue things are. And the grue things are more alike than some other even more disjunctive bunches. So as well as positing perfectly natural properties, Lewis posits a relation of more and less natural on properties. He suggests that we just need to take the perfectly natural as primitive, and we can define the naturalness of other properties in terms of it. The idea is that the naturalness of a property is a function of the complexity of that property's definition in terms of perfectly natural properties. It isn't at all obvious that this suggestion will capture the intuitive idea, and Lewis does not defend it at any length.

One of the roles of natural properties is in induction. Other things being equal, more natural properties are more projectible. That's Lewis's solution to Goodman's problem. We don't project grue because doing so would conflict with projecting green, and green is more natural.

Rational agents have beliefs that follow inductively from their evidence. So rational agents tend to have beliefs involving natural rather than unnatural properties. If the contents of beliefs are properties, as we suggested in the previous section, we can simplify this a bit and say

that rational agents have beliefs whose contents are natural properties. Given interpretationism, what's true of rational agents is true, other things equal, of any agent, since the correct interpretation of an agent's beliefs assumes they are rational. So, other things being equal, our beliefs have more rather than less natural content. So, other things being equal, we believe that grass is green not grue. That's Lewis's solution to the Kripkenstein (and Putnam) problems. Even if my dispositions would be consistent with my believing grass is grue, ascribing that to me would uncharitably attribute gratuitous irrationality to me. Since a correct interpretation doesn't ascribe gratuitous irrationality, that ascription would be incorrect. So I don't believe grass is grue.

Natural properties will play a major role for Lewis. We've already seen one place where it turns out they are needed; namely, in saying what it is for two worlds to have an 'exact match' of spatiotemporal regions. What Lewis means by that is that the regions are intrinsic duplicates. And the way he analyses intrinsic duplication in (1983b) is that two things are duplicates if they have the same intrinsic properties. We will see many other uses of natural properties as we go along, particularly in the discussion of Humean supervenience in section 5.

This topic, natural properties, was one of very few topics where Lewis had a serious change of view over the course of his career. Of course, Lewis changed the details of many of his views, in response to criticism and further thought. But the idea that some properties could be natural, could make for objective similarity, in ways that most sets of possibilia do not, is notably absent from his writings before "New Work". Indeed, as late as "Individuation by Acquaintance and by Stipulation" (1983a), he was rather dismissive of the idea. But natural properties came to play central roles in his metaphysics and, as we see here, his theory of mind. As Lewis notes in "New Work", much of the impetus for his change of view came from discussions with D. M. Armstrong, and from the arguments in favour of universals that Armstrong presented in his (1978).

5 Humean Supervenience

Many of David Lewis's papers in metaphysics were devoted to setting out, and defending, a doctrine he called "Humean Supervenience". Here is Lewis's succinct statement of the view.

> It is the doctrine that all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another. (1986c, ix)

The doctrine can be factored into two distinct theses. The first is the thesis that, in John Bigelow's words, "truth supervenes on being". That is, all the truths about a world supervene on the distribution of perfectly natural properties and relations in that world. The second is the thesis that the perfectly natural properties and relations in this world are intrinsic properties of point-sized objects, and spatiotemporal relations. Lewis held that the first of these was necessary and a priori. (See, for instance, "Parts of Classes" (1991), "Reduction of Mind", "Truthmaking and Difference-making" (2001c).) The second is contingently true if true at all. Indeed, modern physics suggests that it is not true (Maudlin, 2007, Ch. 2). Lewis was aware of this. His aim in defending Humean supervenience was to defend, as he put it, its "tenability" (1986b, xi). We will return at the end of this section to the question of why he might have wanted to do this. For now, we will focus on how he went about this project.

The primary challenge to Humean supervenience comes from those who hold that providing a subvenient basis for all the truths of this world requires more than intrinsic properties of point-sized objects and spatiotemporal relations. Some of these challenges come from theorists who think best physics will need non-spatiotemporal relations in order to explain Bell's Theorem. But more commonly it comes from those who think that grounding the modal, the nomic or the mental requires adding properties and relations to any Humean mosaic constructed from properties found in fundamental physics. (I'm using 'mental' here to cover all the properties that Lewis considered mental, broadly construed. This includes contents, since Lewis thought content was grounded in mental content, and value, since he thought values were grounded in idealised desires. So it's a fairly broad category, and there is a lot that isn't obviously reducible to fundamental physics. As we'll see, Lewis attempts to reduce it all step-by-step.)

We've discussed in the previous section how Lewis aimed to reduce the mental to the nomic. (Or at least much of it; we'll return to the question of value in section 7.5.) We'll discuss in the next section his distinctive modal metaphysics. In this section we'll look at how he attempted to locate the nomic in the Humean mosaic. Lewis's aim was to show that nomic properties and relations could be located in the Humean mosaic by locating them as precisely and as explicitly as he could. So the location project revealed a lot about these nomic features. We'll spend the next two subsections looking at the two important parts of this project. Notably, they are two parts where Lewis refined his views several times on the details of the location.

5.1 Laws and Chances

Lewis's reductionist project starts with laws of nature. Building on some scattered remarks by Ramsey and Mill, Lewis proposed a version of the 'best-system' theory of laws of nature. There is no paper devoted to this view, but it is discussed in section 3.3 of *Counterfactuals*, in "New Work For a Theory of Universals", extensively in Postscript C to the reprint of "A Subjectivist's Guide to Objective Chance" in (1986c), and in "Humean Supervenience Debugged" (1994a).

The simple version of the theory is that the laws are the winners of a 'competition' among all collections of truths. Some truths are simple, e.g. the truth that this table is brown. Some truths are strong; they tell us a lot about the world. For example, the conjunction of every truth in this Encyclopedia rules out a large chunk of modal space. Typically, these are exclusive categories; simple truths are not strong, and strong truths are not simple. But there are some exceptions. The truth that any two objects are attracted to one another, with a force proportional to the product of their masses and inversely proportional to the distance between them, is relatively simple, but also quite strong in that it tells us a lot about the forces between many distinct objects. The laws, says Lewis, are these simple but strong truths.

Two qualifications are needed before we get to Lewis's 1973 view of laws. It is collections of truths, not individual truths, that are measured and compared for simplicity and strength. And it is not every truth in the winning collection (or best system), but only the generalisations within it, that are laws. So even if the best system includes particular facts about the Big Bang or its immediate aftermath, e.g. that the early universe was a low entropy state, those facts are not laws on Lewis's view.

In "New Work For a Theory of Universals", Lewis notes another restriction that is needed. If we measure the simplicity of some truths by the length of their statement in an arbitrarily chosen language, then any truth at all can be made simple. Let Fx be true iff x is in a world where every truth in this Encyclopedia is true. Then *Everything is F* is simply stateable in

a language containing *F*, and is presumably strong. So *Everything is F* will be a law. But this kind of construction would clearly trivialise the theory of laws. Lewis's solution is to say that we measure the simplicity of a claim by how easily stateable it is in a language where all predicates denote *perfectly natural* properties. He notes that this move requires that the natural properties are specified prior to specifying the laws, which means that we can't reductively specify naturalness in terms of laws. (In any case, since Lewis holds that laws are contingent (1986b, 91) but which properties are natural is not contingent (1986b, 60n), this approach would not be open to Lewis.)

In "Humean Supervenience Debugged", Lewis notes how to extend this theory to indeterministic worlds. Some laws don't say what will happen, but what will have a chance of happening. If the chances of events could be determined antecedently to the laws being determined, we could let facts about chances be treated more or less like any other fact for the purposes of our 'competition'. But, as we'll see, Lewis doesn't think the prospects for doing this are very promising. So instead he aims to reduce laws and chances simultaneously to distributions of properties.

Instead of ranking collections of truths by two measures, strength and simplicity, we will rank them by three, strength, simplicity and fit. A collection of truths that entails that what does happen has (at earlier times) a higher chance of happening has better fit than a collection that entails that what happens had a lower chance of happening. The laws are those generalisations in the collection of truths that do the best by these three measures of strength, simplicity and fit. The collection will entail various 'history-to-chance' conditionals. These are conditionals of the form *If* H_t *then* $P_t(A) = x$, where H_t is a proposition about the history of the world to t, and P_t is the function from propositions to their chance at t. The chance of A at t in w is x iff there is some such conditional *If* H_t *then* $P_t(A) = x$, where H_t is the history of w to t.

The position that I've sketched here is the position that Lewis says that he originally was drawn towards in 1975, and that he endorsed in print in 1994. (The dates are from his own description of the evolution of his views in (1994a).) But in between, in both (1980b) and Postscript C to its reprinting in (1986c), he rejected this position because he thought it conflicted with a non-negotiable conceptual truth about chance. This truth was what he called the "Principal Principle".

The Principal Principle says that a rational agent conforms their credences to the chances. More precisely, it says the following is true. Assume we have a number x, proposition A, time t, rational agent whose evidence is entirely about times up to and including t, and a proposition E that (a) is about times up to and including t and (b) entails that the chance of A at t is x. In any such case, the agent's credence in A given E is x.

An agent who knows what happens after t need not be guided by chances at t. If I've seen the coin land heads, that its chance of landing heads was 0.5 at some earlier time is no reason to have my credence in heads be 0.5. Conversely, if all I know is that the chance is 0.5, that's no reason for my conditional credence in heads to be 0.5 conditional on anything at all. Conditional on it landing heads, my credence in heads is 1, for instance. But given these two restrictions, the Principal Principle seems like a good constraint. Lewis calls evidence about times after t 'inadmissible', which lets us give a slightly more concise summary of what the Principal Principle says. For agents with no inadmissible evidence, the rational credence in A, conditional on the chance of A being x, combined with any admissible evidence, is x.

The problem Lewis faced in the 1980s papers is that the best systems account of chance makes the Principal Principle either useless or false. Here is a somewhat stylised example. (I make no claims about the physical plausibility of this setup; more plausible examples would be more complicated, but would make much the same point.) Let t be some time before any particle has decayed. Let A be the proposition that every radioactive particle will decay before it reaches its actual half-life. At t, A has a positive chance of occurring. Indeed, its chance is 1 in 2^n, where n is the number of radioactive particles in the world. (Assume, again for the sake of our stylised example, that n is finite.) But if A occurred, the best system of the world would be different from how it actually is. It would improve fit, for instance, to say that the chance of decay within the actual half-life would be 1. So someone who knows that the chance of A is 1 in 2^n knows that A won't happen.

Lewis called A an 'undermining' future; it has a chance of happening, but if it happens the chances are different. The problem with underminers is that they conflict with the Principal Principle. Someone who knows the chance of A should, by the Principal Principle, have credence 1 in 2^n that A will happen. But given the chance of A, it is possible to deduce A, and hence have credence in A. This looks like an inconsistency, so like any principle that implies a contradiction, the Principal Principle must be false. The most obvious way out is to say that information about the chance of A is inadmissible, since it reveals something about the future, namely that A doesn't occur. But to say that chances are inadmissible is to make the Principal Principle useless. So given the best systems theory of laws and chances, the Principal Principle is either false or useless. Since the Principal Principle is neither false nor useless, Lewis concluded in these 1980s papers that the best systems theory of laws and chances was false.

The problem with this was that it wasn't clear what could replace the best systems theory. Lewis floated two approaches in the postscripts to the reprinting of (1980b), one based on primitive chances, and the other based on history-to-chance conditionals being necessary. But neither seemed metaphysically plausible, and although each was consistent with the Principal Principle, they made it either mysterious (in the first case) or implausible (in the second). A better response, as set out in "Humean Supervenience Debugged", was to qualify the Principal Principle. Lewis said that what was really true was the "New Principle". His proposal was based on ideas developed by Ned Hall (1994) and Michael Thau (1994).

We'll explain the New Principle by starting with a special case of the old Principle. Let T be the 'theory of chance' for the world, the conjunction of all history-to-chance conditionals. And let H be the history of the world to t. Assuming T is admissible, the old Principal Principle says that the credence in A given $H \wedge T$ should be the chance of A at t. The New Principle says that the credence in A given $H \wedge T$ should be the chance of A given T at t. That is, where C is the agent's credence function, and P is the chance function, and the agent has no inadmissible evidence, it should be that $C(A \mid H \wedge T) = P(A \mid T)$. This compares to the old principle, which held that $C(A \mid H \wedge T) = P(A)$.

That's the special case of the New Principle for an agent with no inadmissible evidence. The general case follows from this special case. In general, assuming the agent has no inadmissible evidence, the rational credence in A given E is the expected value, given E, of the chance of A given $H \wedge T$. That is, where C is the agent's credence function, and P is the chance function, it should be the sum across all possible combinations of H and T of $C(H \wedge T \mid E)P(A \mid H \wedge T)$.

The New Principle is, Lewis argues, consistent with the best systems theory of laws and chances. Lewis had originally thought that any specification of chance had to be consistent with the Principal Principle. But in later works he argued that the New Principle was a close enough approximation to the Principal Principle that a theory of chances consistent with it was close enough to our pre-theoretic notion of chance to deserve the name. So he could, and did, happily endorse the best systems theory of laws and chance.

5.2 Causation

In "Causation" (1973a), Lewis put forward an analysis of causation in terms of counterfactual dependence. The idea was that event *B* was counterfactually dependent on event *A* if and only if the counterfactual *Had A not occurred, B would not have occurred* was true. Then event *C* causes event *E* if and only if there is a chain *C*, D_1, ..., D_n, *E* such that each member in the chain (except *C*) is counterfactually dependent on the event before it. In summary, causation is the ancestral of counterfactual dependence.

The reasoning about chains helped Lewis sidestep a problem that many thought unavoidable for a counterfactual theory of causation, namely the problem of pre-empting causes. Imagine that Suzy throws a rock, the rock hits a window and the window shatters. Suzy's throw caused the window to shatter. But there is a backup thrower—Billy. Had Suzy not thrown, Billy would have thrown another rock and broken the window. So the window breaking is not counterfactually dependent on Suzy's throw. Lewis's solution was to posit an event of the rock flying towards the window. Had Suzy not thrown, the rock would not have been flying towards the window. And had the rock not been flying towards the window, the window would have not shattered. Lewis's thought here is that it is Suzy's throwing that causes Billy to not throw; once she has thrown Billy is out of the picture and the window's shattering depends only on what Suzy's rock does. So we avoid this problem of pre-empters.

Much of the argumentation in "Causation" concerns the superiority of the counterfactual analysis to deductive-nomological theories. These arguments were so successful that from a contemporary perspective they seem somewhat quaint. There are so few supporters of deductive-nomological theories in contemporary metaphysics that a modern paper would not spend nearly so much time on them.

After "Causation" the focus, at least of those interested in reductive theories, moved to counterfactual theories. And it became clear that Lewis had a bit of work left to do. He needed to say more about the details of the notion of counterfactual dependence. He did this in "Counterfactual Dependence and Time's Arrow" (1979b), as discussed in section 2. He needed to say more about the nature of events. In "Events" (1986a) he said that they were natural properties of regions of space-time. And prodded by Jaegwon Kim (1973), he needed to add that *A* and *B* had to be wholly distinct events for *B* to counterfactually depend on *A*. The alternative would be to say that an event's happening is caused by any essential part of the event, which is absurd.

But the biggest problem concerned what became known as "late pre-emption". In the rock throwing example above, we assumed that Billy decided not to throw when he saw Suzy throwing. But we can imagine a variant of the case where Billy waits to see whether Suzy's rock hits, and only then decides not to throw. In such a case, it is the window's shattering, not anything prior to this, that causes Billy not to throw. That means that there is no event

between Suzy's throw and the window's shattering on which the shattering is counterfactually dependent.

Lewis addressed this issue in "Redundant Causation", one of the six postscripts to the reprinting of "Causation" in (1986c). He started by introducing a new concept: quasi-dependence. *B* quasi-depends on *A* iff there is a process starting with *A**, and ending with *B** , and *B** counterfactually depends on *A**, and the process from *A** to *B** is an intrinsic duplicate of the process from *A* to B, and the laws governing the process from *A** to *B** (i.e. the laws of the world in which *A** and *B** happen) are the same as the laws governing the process from *A* to *B*. In short, quasi-dependence is the relation you get if you start with dependence, then add all of the duplicates of dependent processes. Causation is then the ancestral of quasi-dependence. Although the window's shattering does not depend on Suzy's throw, it does quasi-depend on it. That's because there is a world, with the same laws, with a duplicate of Suzy's throw, but Billy determined not to throw, and in that world the window shatters in just the same way, and depends on Suzy's throw.

Eventually, Lewis became unsatisfied with the quasi-dependence based theory. In "Causation as Influence" (2000; 2004a) he set out several reasons for being unhappy with it, and a new theory to supersede it.

One argument against it is that it makes causation intrinsic to the pair *C* and *E*, but some cases, especially cases of *double prevention*, show that causation is extrinsic. Double prevention occurs when an event, call it *C*, prevents something that would have prevented *E* from happening. Intuitively, these are cases of causation. Indeed, when we look at the details we find that many everyday cases of causation have this pattern. But that *C* causes *E* does not depend on the intrinsic natures of *C* and *E*. Rather, it depends on there being some threat to *E*, a threat that *C* prevents, and the existence of threats is typically extrinsic to events.

Another argument is that quasi-dependence cannot account for what came to be known as 'trumping pre-emption'. Lewis illustrated this idea with an example from Jonathan Schaffer (2000). The troops are disposed to obey all orders from either the Sergeant or the Major. But they give priority to the Major's orders, due to the Major's higher rank. Both the Major and the Sergeant order the troops to advance, and they do advance. Intuitively, it is the Major, not the Sergeant, who caused the advance, since the Major's orders have priority. But the advance does quasi-depend on the Sergeant's orders, since in a world where the Major doesn't make an order, the advance does depend on the Sergeant.

Lewis's alternative theory relied on changing the definition of counterfactual dependence. The theory in "Causation" was based on what he came to call 'whether-whether' dependence. What's crucial is that *whether B* happens depends counterfactually on *whether A* happens. The new theory was based on what we might call 'how-how' dependence. Lewis says that *B* depends on *A* if there are large families of counterfactuals of the form *If A had happened in this way, then B would have happened in that way*, and the ways in which *B* would happen are systematically dependent on the ways in which *A* happens. How much *A* influences *B* depends on how big this family is, how much variation there is in the way *B* changes, and how systematic the influence of *A* on *B* is. He then defines causation as the ancestral of this notion of counterfactual dependence.

On this new theory, causation is a degree concept, rather than an 'all-or-nothing' concept, since counterfactual dependence comes in degrees. Sometimes Lewis says we properly ignore small amounts of causation. For instance, the location of nearby parked cars influences the

smashing of a window by a rock in virtue of small gravitational effects of the cars on the flight of the rock. But it's very little influence, and we properly ignore it most of the time.

There are two other notable features of "Causation as Influence". It contains Lewis's most comprehensive defence of the transitivity of causation. This principle was central to Lewis's theory of causation from the earliest days, but had come under sustained attack over the years. And the paper has a brief attack on non-Humean theories that take causation to be a primitive. Lewis says that these theories can't explain the variety of causal relations that we perceive and can think about. These passages mark an interesting change in what Lewis took to be the primary alternatives to his counterfactuals based reductionism. In 1973 the opponents were other kinds of reductionists; in 2000 they were the non-reductionists.

5.3 Why Humean Supervenience

Given these concepts, a number of other concepts fall into place. Dispositions are reduced to counterfactual dependencies, though as is made clear in "Finkish Dispositions" (1997b), the reduction is not as simple as it might have seemed. Perception is reduced to dispositions and causes. (See, for instance, "Veridical Hallucination and Prosthetic Vision" (1980c).) We discussed the reduction of mental content to dispositions and causes in section 4. And we discussed the reduction of linguistic content to mental content in section 1. Values are reduced to mental states in "Dispositional Theories of Value" (1989b).

But we might worry about the very foundation of the project. We started with the assumption that our subvenient base consists of intrinsic properties of point-sized objects and spatiotemporal relations. But Bell's inequality suggests that modern physics requires, as primitive, other relations between objects. (Or it requires intrinsic properties of dispersed objects.) So Humean supervenience fails in this world.

Lewis's response is somewhat disarming. Writing in 1986, part of his response is scepticism about the state of quantum mechanics. (There is notably less scepticism in "How Many Lives Has Schr odinger's Cat" (2004b).) But the larger part of his response is to suggest that scientific challenges to Humean supervenience are outside his responsibility.

> Really, what I uphold is not so much the truth of Humean supervenience as the *tenability* of it. If physics itself were to teach me that it is false, I wouldn't grieve ... What I want to fight are *philosophical* arguments against Humean supervenience. When philosophers claim that one or another common-place feature of the world cannot supervene on the arrangement of qualities, I make it my business to resist. Being a commonsensical fellow (except where unactualized possible worlds are concerned) I will seldom deny that the features in question exist. I grant their existence, and do my best to show how they can, after all, supervene on the arrangement of qualities. (1986c, xi)

We might wonder why Lewis found this such an *interesting* project. If physics teaches that Humean supervenience is false, why care whether there are also philosophical objections to it? There are two (related) reasons why we might care.

Recall that we said that Humean supervenience is a conjunction of several theses. One of these is a thesis about which perfectly natural properties are instantiated in this world, namely local ones. That thesis is threatened by modern physics. But the rest of the package, arguably,

is not. In particular, the thesis that all facts supervene on the distribution of perfectly natural properties and relations does not appear to be threatened. (Though see (Maudlin, 2007, Ch. 2) for a dissenting view.) Nor is the thesis that perfectly natural properties and relations satisfy a principle of recombination threatened by modern physics. The rough idea of the principle of recombination is that any distribution of perfectly natural properties is possible. This thesis is Lewis's version of the Humean principle that there are no necessary connections between distinct existences, and Lewis is determined to preserve as strong a version of it as he can.

Although physics does not seem to challenge these two theses, several philosophers do challenge them on distinctively philosophical grounds. Some of them suggest that the nomic, the intensional, or the normative do not supervene on the distribution of perfectly natural properties. Others suggest that the nomic, intentional, or normative properties are perfectly natural, and as a consequence perfectly natural properties are not freely recombinable. The philosophical arguments in favour of such positions rarely turn on the precise constitution of the Humean's preferred subvenient base. If Lewis can show that such arguments fail in the setting of classical physics, then he'll have refuted all of the arguments against Humean superveience that don't rely on the details of modern physics. In practice that means he'll have refuted many, though not quite all, of the objections to Humean supervenience.

A broader reason for Lewis to care about Humean supervenience comes from looking at his overall approach to metaphysics. When faced with something metaphysically problematic, say free will, there are three broad approaches. Some philosophers will argue that free will can't be located in a scientific world-view, so it should be eliminated. Call these 'the eliminativists'. Some philosophers will agree that free will can't be located in the scientific world-view, so that's a reason to expand our metaphysical picture to include free will, perhaps as a new primitive. Call these 'the expansionists'. And some philosophers will reject the common assumption of an incompatibility. Instead they will argue that we can have free will without believing in anything that isn't in the scientific picture. Call these 'the compatibilists'.

As the above quote makes clear, Lewis was a compatibilist about most questions in metaphysics. He certainly was one about free will. ("Are We Free to Break the Laws?" (1981a).) And he was a compatibilist about most nomic, intentional and normative concepts. This wasn't because he had a global argument for compatibilism. Indeed, he was an eliminativist about religion ("Anselm and Actuality" (1970a), "Divine Evil" (2007)). And in some sense he was an expansionist about modality. Lewis may have contested this; he thought introducing more worlds did not increase the number of kinds of things in our ontology, because we are already committed to there being at least one world. As (Melia, 1992, 192) points out though, the inhabitants of those worlds include all kinds of things not found in, or reducible to, fundamental physics. They include spirits, gods, trolls and every other consistent beast imaginable. So at least when it came to what there is, as opposed to what there actually is, Lewis's ontology was rather expansionist.

For all that, Lewis's default attitude was to accept that much of our common-sense thinking about the nomic, the intentional and the normative was correct, and that this was perfectly compatible with this world containing nothing more than is found in science, indeed than is found in fundamental physics.

Compatibilists should solve what Frank Jackson calls 'the location problem' (Jackson 1998). If you think that there are, say, beliefs, and you think that having beliefs in one's metaphysics doesn't commit you to having anything in your ontology beyond fundamental physics, then

you should, as Jackson puts it, be able to *locate* beliefs in the world described by fundamental physics. More generally, for whatever you accept, you should be able to locate it in the picture of the world you accept.

This was certainly the methodology that Lewis accepted. And since he thought that so much of our common sense worldview was compatible with fundamental physics, he had many versions of the location problem to solve. One way to go about this would be to find exactly what the correct scientific theory is, and locate all the relevant properties in that picture. But this method has some shortcomings. For one thing, it might mean having to throw out your metaphysical work whenever the scientific theories change. For another, it means having your metaphysics caught up in debates about the best scientific theories, and about their interpretation. So Lewis took a somewhat different approach.

What Lewis's defence of Humean supervenience gives us is a recipe for locating the nomic, intentional and normative properties in a physical world. And it is a recipe that uses remarkably few ingredients; just intrinsic properties of point-sized objects, and spatio-temporal relations. It is likely that ideal physics will have more in it than that. For instance, it might have entanglement relations, as are needed to explain Bell's inequality. But it is unlikely to have less. And the more there is in fundamental physics, the *easier* it is to solve the location problem, because the would-be locator has more resources to work with.

The upshot of all this is that a philosophical defence of Humean supervenience, especially a defence like Lewis's that shows us explicitly how to locate various folk properties in classical physics, is likely to show us how to locate those properties in more up-to-date physics. So Lewis's defence of Humean supervenience then generalises into a defence of the compatibility of large swathes of folk theory with ideal physics. And the defence is consistent with the realist principle that truth supervenes on being, and with the Humean denial of necessary connections between distinct existences. And that, quite clearly, is a philosophically interesting project.

6 Modal Realism

This entry has been stressing Lewis's many and diverse contributions to philosophy. But there is one thesis with which he is associated above all others: modal realism. Lewis held that this world was just one among many like it. A proposition, p is possibly true if and only if p is true in one of these worlds. Relatedly, he held that individuals like you or I (or this computer) only exist in one possible world. So what it is for a proposition like *You are happy* to be true in another world is not for you to be happy in that world; you aren't in that world. Rather, it is for your *counterpart* to be happy in that world.

Lewis wrote about modal realism in many places. As early as *Counterfactuals* he wrote this famous passage.

> I believe, and so do you, that things could have been different in countless ways. But what does this mean? Ordinary language permits the paraphrase: there are many ways things could have been besides the way they actually are. I believe that things could have been different in countless ways; I believe permissible paraphrases of what I believe; taking the paraphrase at its face value, I therefore believe in the existence of entities that might be called 'ways things could have been.' I prefer to call them 'possible worlds.' (1973b, 84)

And Lewis used counterpart theory throughout his career to resolve metaphysical puzzles in fields stretching from personal identity ("Counterparts of Persons and Their Bodies" (1971)) to truthmaker theory ("Things qua Truthmakers" (2003)). Indeed, Lewis's original statement of counterpart theory is in one of his first published metaphysics papers ("Counterpart Theory and Quantified Modal Logic" (1968)).

But the canonical statement and defence of both modal realism and counterpart theory is in *On the Plurality of Worlds* (1986b), the book that grew out of his 1984 John Locke lectures. This section will follow the structure of that book.

The little 'argument by paraphrase' from *Counterfactuals* is a long way from an argument for Lewis's form of modal realism. For one thing, the argument relies on taking a folksy paraphrase as metaphysically revealing; perhaps we would be better off treating this as just a careless manner of speaking. For another, the folksy paraphrase Lewis uses isn't obviously innocuous; like many other abstraction principles it could be hiding a contradiction. And the argument does little to show that other possible worlds are concreta; talking of them as ways things could be makes them sound like properties, which are arguably abstracta if they exist at all. The first three chapters of *Plurality* address these three issues. The fourth chapter is an extended discussion of the place of individuals in modal realism. We'll look at these chapters in order.

6.1 A Philosophers' Paradise

The short argument from *Counterfactuals* that I quoted seems deeply unQuinean. Rather than saying that possible worlds exist because they are quantified over in the best paraphrase of our theories, Lewis says they exist because they are quantified over in just one paraphrase of our theories. To be sure, he says this is a permissible paraphrase. On the other hand, there is vanishingly little defence of its permissibility.

In the first chapter of *Plurality* Lewis takes a much more Quinean orthodox line. He argues, at great length, that the best version of many philosophical theories requires quantification over possibilities. In traditional terms, he offers an extended indispensibility argument for unactualised possibilities. But traditional terms are perhaps misleading here. Lewis does not say that possibilities are absolutely indispensible, only that they make our philosophical theories so much better that we have sufficient reason to accept them.

There are four areas in which Lewis thinks that possible worlds earn their keep.

Modality Traditional treatments of modal talk in terms of operators face several difficulties. They can't, at least without significant cost, properly analyse talk about contingent existence, or talk about modal comparatives, or modal supervenience theses. All of these are easy to understand in terms of quantification across possibilities.

Closeness Our best theory of counterfactuals, Lewis's theory, relies on comparisons between possible worlds. Indeed, it relies on comparisons between this world and other worlds. Such talk will be hard to paraphrase away if worlds aren't real.

Content Lewis argues, in part following Stalnaker (1984), that our best theory of mental and verbal content analyses content in terms of sets of possibilities. This, in turn, requires that the possibilities exist.

Properties We often appear to quantify over properties. The modal realist can take properties to be sets of possibilia, and take such quantification at face value. In his discussion of properties here, Lewis expands upon his theory of natural properties that he introduced in "New Work for a Theory of Universals", and that we discussed in section 3.

After arguing that we are best off in all these areas of philosophy if we accept unactualised possibilities, Lewis spends the rest of chapter 1 saying what possible worlds are on his view. He isn't yet arguing for this way of thinking about possible worlds; that will come in chapter 3. For now he is just describing what he takes to be the best theory of possible worlds. He holds that possible worlds are isolated; no part of one is spatio-temporally related to any other world. Indeed, he holds that lack of spatio-temporal relation (or something like it) is what marks individuals as being in different worlds. So his theory has the somewhat odd consequence that there could not have been two parts of the world that aren't spatio-temporally connected. He holds that worlds are concrete, though spelling out just what the abstract/concrete distinction comes to in this context isn't a trivial task. And he holds that worlds are plenitudinous. There is a world for every way things could be. And worlds satisfy a principle of recombination: shape and size permitting, any number of duplicates of any number of possible things can co-exist or fail to co-exist.

6.2 Paradox in Paradise?

Chapter 2 deals with several objections to modal realism. Some of these objections claim that modal realism leads to paradox. Other objections claim that it undermines our ordinary practice. We will look at two examples of each.

Peter Forrest and D. M. Armstrong (1984) argue that modal realism leads to problems given the principle of recombination. An unrestricted principle of recombination says that for any things that could exist, there is a world in which there is a duplicate of all of them. Forrest and Armstrong apply the principle by taking the things to be the different possible worlds. A world containing a duplicate of all the worlds would, they show, be bigger than any world. But by the principle it would also be a world. Contradiction. Lewis' reply is to deny the unrestricted version of the principle. He insists that there is independent reason to qualify the principle to those things whose size and shape permits them to be fit into a single world. Without an unrestricted principle of recombination, there is no way to create the large world that's at the heart of Forrest and Armstrong's paradox.

David Kaplan argued that there could be no cardinality of the worlds. Kaplan did not publish this argument, so Lewis replies to the version presented by Martin (Davies, 1981, 262). On Lewis's theory, every set of worlds is a proposition. For any proposition, says Kaplan, that proposition might be the only proposition being thought by a person at location l at time t. So for each proposition, there is a world where it (alone) is thought by a person at location l at time t. That means there is a one-one correspondence between the sets of worlds and a subset of the worlds. Contradiction. Lewis's reply is to deny that every proposition can be thought. He claims that functionalism about belief, plus the requirement that beliefs latch onto relatively natural properties, mean that most propositions cannot be thought, and this blocks the paradox.

Peter Forrest (1982) argues that modal realism leads to inductive scepticism. According to modal realism, there are other thinkers very much like us who are deceived by their surroundings. Given this, we should doubt our inductive inferences. Lewis's reply is that modal realism does not make inductive challenges any worse than they were before. It is common ground that inductive inference is fallible. That is, it is common ground that these inferences could fail. Thinking of the possibilities of failure as concrete individuals might focus the mind on them, and hence make us less confident, but does not seem to change the inference's justificatory status. Lewis's argument seems hard to dispute here. Given the mutually agreed upon fact that the inference could fail, it's hard to see what epistemological cost is incurred by agreeing that it does fail for someone kind of like the inferrer in a distant possible world.

Robert Adams (1974) argues that modal realism leads to surprising results in moral philosophy. The modal realist says that the way things are, in the broadest possible sense, is not a contingent matter, since we can't change the nature of the pluriverse. Hence we cannot do anything about it. So if moral requirements flow from a requirement to improve the way things are, in this broadest possible sense, then there are no moral requirements. Lewis rejects the antecedent of this conditional as something that only an extreme utilitarian could accept. What is crucial about morality is that *we* not do evil. Even if their actions won't make a difference to the nature of the pluriverse, a virtuous agent will not want to, for instance, cause suffering. By rejecting the view that in our moral deliberations we should care about everyone, possible and actual, equally, Lewis avoids the problem.

6.3 Paradise on the Cheap?

In chapter 3 Lewis looks at the alternatives to his kind of modal realism. He takes himself to have established that we need to have possible worlds of some kind in our ontology, but not that these possible worlds must be concrete. In particular, they can be abstract, or what he calls "ersatz" possible worlds. Lewis does not have a single knock-down argument against all forms of ersatzism. Instead he divides the space of possible ersatzist positions into three, and launches different attacks against different ones.

Lewis starts with what he calls "linguistic ersatzism". This is the view that ersatz possible worlds are representations, and the way they represent possibilities is something like the way that language represents possibilities. In particular, they represent possibilities without resembling possibilities, but instead in virtue of structural features of the representation.

He levels three main objections to linguistic ersatzism. First, it takes modality as a primitive, rather than reducing modality to something simpler (like concrete possible worlds). Second, it can't distinguish qualitatively similar individuals in other possible worlds. Lewis argues that will mean that we can't always quantify over possibilia, as we can in his theory. Third, it can't allow as full a range of 'alien', i.e. uninstantiated, natural properties as we would like. Sider (2002) has replied that some of these challenges can be met, or at least reduced in intensity, if we take the pluriverse (i.e. the plurality of worlds) to be what is represented, rather than the individual worlds.

The second theory he considers is what he calls "pictoral ersatzism". This is the view that ersatz possible worlds are representations, and the way they represent possibilities is something like the way that pictures or models represent possibilities. That is, they represent by being similar, in a crucial respect, to what they are representing. The pictoral ersatzist, says Lewis, is caught in something of a bind. If the representations are not detailed enough, they will not

give us enough possibilities to do the job that possible worlds need to do. If they are detailed enough to do that job, and they represent by resembling possibilities, then arguably they will contain as much problematic ontology as Lewisian concrete possible worlds. So they have the costs of Lewis's theory without any obvious advantage.

The final theory he considers is what he calls "magical ersatzism". Unlike the previous two theories, this theory is defined negatively. The magical ersatzist is defined by their denial that possible worlds represent, or at least that they represent in either of the two ways (linguistic and pictoral) that we are familiar with. And Lewis's primary complaint is that this kind of theory is mysterious, and that it could only seem attractive if it hides from view the parts of the theory that are doing the philosophical work. Lewis argues that as soon as we ask simple questions about the relationship that holds between a possibility and actuality if that possibility is actualised, such as whether this is an internal or external relation, we find the magical ersatzist saying things that are either implausible or mysterious.

It isn't clear just who is a magical ersatzist. Lewis wrote that at the time he wrote *Plurality* no one explicitly endorsed this theory. This was perhaps unfair to various *primitivists* about modality, such as Adams (1974), Plantinga (1974) and Stalnaker (1976). Given the negative definition of magical ersatzism, and given the fact that primitivists do not think that possible worlds represent possibilities via any familiar mechanism, it seems the primitivists should count as magical ersatzists, or, as Lewis calls them, "magicians". In any case, if magical ersatzism, in all its varieties, is objectionably mysterious, that suggests ersatzism is in trouble, and hence if we want the benefits of possible worlds, we have to pay for them by accepting concrete possible worlds.

6.4 Counterparts or Double Lives?

The last chapter of *Plurality* changes tack somewhat. Instead of focussing on different ways the world could be, Lewis's focus becomes different ways things could be. The chapter defends, and expands upon, Lewis's counterpart theory.

Counterpart theory was first introduced by Lewis in "Counterpart Theory and Quantified Modal Logic" (1968) as a way of making modal discourse extensional. Instead of worrying just what a name inside the scope of a modal operator might mean, we translate the language of quantified modal logic into a language without operators, but with quantifiers over worlds and other non-actual individuals. So instead of saying $\Box Fa$, we say $\forall w \forall x ((Ww \wedge Ixw \wedge Cxa) \supset Fx)$. That is, for all w and x, if w is a world, and x is in w, and x is a counterpart of a, then Fx. Or, more intuitively, all of a's counterparts are F. The paper shows how we can extend this intuitive idea into a complete translation from the language of quantfied modal logic to the language of counterpart theory. In "Tensions" (1974b) Lewis retracts the claim that it is an advantage of counterpart theory over quantified modal logic that it is extensional rather than intensional, largely because he finds the distinction between these two notions much more elusive than he had thought. But he still thought counterpart theory had a lot of advantages, and these were pressed in chapter 4.

The intuitive idea behind counterpart theory was that individuals, at least ordinary individuals of the kind we regularly talk about, are world-bound. That is, they exist in only one world. But they do not have all of their properties essentially. We can truly say of a non-contender, say Malloy, that he could have been a contender. In the language of possible worlds, there is a possible world w such that, according to it, Malloy is a contender. But what in turn does

this mean? Does it mean that Malloy himself is in *w*? Not really, according to counterpart theory. Rather, a counterpart of Malloy's is a contender in *w*. And Malloy himself has the modal property *could have been a contender* in virtue of having a counterpart in *w* who is a contender. This way of thinking about modal properties of individuals has, claims Lewis, a number of advantages.

For one thing, it avoids an odd kind of inconsistency. Malloy might not only have been a contender, he might have been 6 inches taller. If we think that is because there is a world in which Malloy himself is 6 inches taller, then it seems like we're saying that Malloy can have two heights, his actual height and one 6 inches taller. And that looks inconsistent. The obvious way out of this is to say that he bears one height in relation to this world, and another to another world. But that turns height from an intrinsic property into a relation, and that seems like a mistake. Lewis thinks this problem, what he dubs the 'problem of accidental intrinsics', is a reason to deny that Malloy himself is in multiple worlds.

For another, it allows us a kind of inconstancy in our modal predications. Could Malloy have been brought by a stork, or must he have had the parents he actually had? In some moods we think one, in other moods we think another. Lewis thinks that counterpart theory can reflect our indecision. There is a world with someone brought by a stork who has a life much like Malloy's. Is he one of Malloy's counterparts? Well, he is according to some counterpart relations, and not according to others. When one of the former relations is contextually salient, it's true to say that Malloy could have been brought by a stork. When more demanding counterpart relations are salient, he isn't one of Malloy's counterparts, and indeed all of Malloy's counterparts share his parents. (More precisely, all of his counterparts have parents who are counterparts of Malloy's actual parents.) In those contexts, it is true to say that one's parentage is essential. Throughout his career, Lewis uses this inconstancy of the counterpart relation to resolve all manner of metaphysical puzzles, from puzzles about personal identity (1971) to puzzles about truthmakers (2003). The final section of *Plurality* is Lewis's most extended argument that this variability of the counterpart relation is a strength, not a weakness, of the theory.

7 Other Writings

Lewis wrote a lot that isn't covered by the broad categories we've discussed so far. The point of this section is to provide a sample of that material. It isn't close to being comprehensive. It doesn't include his treatment of qualia in (1988d) and (1995). It doesn't include his contributions to causal decision theory in (1979d) and (1981b). It goes very quickly over his many papers in ethics. And it skips his contributions to debates about non-classical logics, such as (1982) and (1990). We've tried to restrict attention to those areas where Lewis's contributions were groundbreaking, influential, and set out a new positive theory. Shockingly, there is a lot to cover that meets those constraints, and is not included in the above survey of the major themes of his philosophy.

7.1 Mathematics and Mereology

Parts of Classes (1991) and "Mathematics is Megethology" (1993b) consider the distinctive philosophical problems raised by set theory. As Lewis notes, it is widely held that all of mathematics reduces to set theory. But there is little consensus about what the metaphysics of set theory is. Lewis puts forward two proposals that might, collectively, help to clarify matters.

The first proposal is what he calls the *Main Thesis*: "The parts of a class are all and only its subclasses" (1991, 7). By 'class' here, Lewis does not mean 'set'. Classes are things with members. Some classes are proper classes, and hence not sets. And one set, the null set, has no members, so is not a class. Individuals, for Lewis, are things without members. Since the null set has no members, it is an individual. But the overlap between the sets and the classes is large; most sets we think about are classes.

The big payoff of the Main Thesis is that it reduces the mysteries of set theory to a single mystery. Any class is a fusion of singletons, i.e., sets with one member. If we understand what a singleton is, and we understand what fusions are, then we understand all there is to know about classes, and about sets. That's because any set is just the fusion of the singletons of its members.

But singletons are deeply mysterious. The usual metaphors that are used to introduce sets, metaphors about combining or collecting or gathering multiple things into one are less than useless when it comes to understanding the relationship between a singleton and its member. In (1993b), Lewis settles for a structuralist understanding of singletons. He also says that he "argued (somewhat reluctantly) for a structuralist' approach to the theory of singleton functions" in (1991), though on page 54 of (1991) he appears to offer qualified resistance to structuralism.

One of the technical advances of (1991) and (1993b) was that they showed how a structuralist account of set theory was even possible. This part of the work was co-authored with John P. Burgess and A. P. Hazen. Given a large enough universe (i.e., that the cardinality of the mereological atoms is an inaccessible cardinal), and given plural quantification, we can say exactly what constraints a function must satisfy for it to do the work we want the singleton function to do. (By 'the singleton function' I mean the function that maps anything that has a singleton onto its singleton. Since proper classes don't have singletons, and nor do fusions of sets and objects, this will be a partial function.) Given that, we can understand mathematical claims made in terms of sets/classes as quantifications over singleton functions. That is, we can understand any claim that would previous have used 'the' singleton function as a claim of the form *for all s*: ...*s*...*s*..., where the terms *s* go where we would previously have referred to 'the' singleton function. It is provable that this translation won't introduce any inconsistency into mathematics (since there are values for *s*), or any indeterminacy (since the embedded sentence ...*s*...*s*... has the same truth value for any eligible value for *s*).

Should we then adopt this structuralist account, and say that we have removed the mysteries of mathematics? As noted above, Lewis is uncharacteristically equivocal on this point, and seemed to change his mind about whether structuralism was, all things considered, a good or a bad deal. His equivocation comes from two sources. One worry is that when we work through the details, some of the mysteries of set theory seem to have been relocated rather than solved. For instance, if we antecedently understood the singleton function, we might have thought it could be used to explain why the set theoretic universe is so large. Now we have to simply posit a very large universe. Another is that the proposal is in some way revisionary, since it takes ordinary mathematical talk to be surreptitiously quantificational. *Parts of Classes* contains some famous invective directed against philosophers who seek to overturn established science on philosophical grounds.

> I'm moved to laughter at the thought of how *presumptous* it would be to reject mathematics for philosophical reasons. How would *you* like the job of telling the mathematicians that they must change their ways, and abjure countless errors, now that *philosophy* has discovered that there are no classes? Can you tell them, with a straight face, to follow philosophical argument wherever it may lead? If they challenge your credentials, will you boast of philosophy's other great discoveries: that motion is impossible, that a Being than which no greater can be conceived cannot be conceived not to exist, that it is unthinkable that anything exists outside the mind, that time is unreal, that no theory has ever been made at all probable by evidence (but on the other hand that an empirically adequate ideal theory cannot possibly be false), that it is a wide-open scientific question whether anyone has ever believed anything, and so on, and on, *ad nauseum*? Not me! (1991, 59)

And yet Lewis's positive theory here is somewhat revisionary. It doesn't revise the truth value of any mathematical claim, but it does revise the understanding of them. Is even this too much revision to make on philosophical grounds? Perhaps not, but it is worrying enough for Lewis to conclude merely that the theory he proposes seems better than the alternatives, not that there is a compelling positive case for its truth.

7.2 Philosophy of Language

Lewis's major contribution to formal semantics was his theory of counterfactual conditionals. But there were several other contributions that he made, both on specific topics in formal semantics, and on the role of semantic theory.

In "Adverbs of Quantification" (1975a), Lewis notes several difficulties in translating sentences involving "usually", "frequently", "rarely" or related adverbs into first-order logic or some similar formal system. Lewis's solution to the puzzles raised involves two formal advances. First, he treats the adverbs as *unselective quantifiers*, binding all free variables in their scope. The second advance concerns the if-clauses in sentences like *Usually, if a team plays well, they win*. It is difficult for various reasons to take the structure of this sentence to involve a quantifier over a compound sentence with a conditional connective. Lewis's second advance is to say that these if-clauses are simply domain restrictors. The 'if' is no more a sentential connective than the 'and' in *New York is between Boston and Washington*. Instead, the if-clause restricts what things the quantifier denoted by 'usually' ranges over.

This paper is not widely read by philosophers, but it has been very influential among linguists, especially semanticists. Indeed, its uptake by semanticists has made it the fourth most cited paper of Lewis's on Google Scholar. His most cited paper on Google Scholar is also in philosophy of language; it is "Scorekeeping in a Language Game" (1979e).

That paper is about conversational dynamics. Lewis develops an extended analogy between the role of context in a conversation and the role of score in a baseball game. One central role of the score is to keep a record of what has already happened. In that way, score is influenced by what happens on the field, or in the conversation. But the causal influence runs in the other way as well. Some events on the field are influenced by the score. You're only out after the *third* strike, for example. Similarly, Lewis holds that context (or the conversational score) can influence, or even be partially constitutive of, what happens in the conversation. If I say "None of the cats are afraid of Barney", which cats I've managed to talk about depends on which cats

are conversationally salient. And in saying this, I've made Barney salient, so the score changes in that respect. That change matters; now I can denote Barney by "he".

Lewis argues that this model can make sense of a number of otherwise puzzling features of language. One notable example of this involves quantification. Most quantifiers we use do not range over the entire universe. We quantify only over a restricted range. Lewis says that it is the salient objects. He also says that this happens not just when we explicitly quantify, but also when we use terms that have a quantificational analysis. He mentions in passing that "knows" might be one such term.

This idea is developed more fully in "Elusive Knowledge" (1996b). Lewis argues that *S knows that p* is true iff *S* is in a position to rule out *all* possibilities in which *p* is false. But when we say *S knows that p*, we don't mean to quantify over all possibilities there are, only over the salient possibilities. The big advantage of Lewis's approach is that it lets him explain the appeal of scepticism. When the sceptic starts talking about fantastic possibilities of error, she makes those possibilities salient. Since we can't rule them out, when we're talking to the sceptic we can't say we know very much. But since those possibilities aren't usually salient, we are usually correct in our knowledge-ascriptions. So Lewis lets the sceptic win any debate they are in, without conceding that ordinary knowledge-ascriptions are false.

The kind of position Lewis defends here, which came to be known as contextualism, has been a central focus of inquiry in epistemology for the last fifteen years. "Elusive Knowledge", along with papers such as Cohen (1986) and DeRose (1995) founded this research program.

7.3 Bayesian Philosophy

This subsection is largely about two pairs of papers: "Probabilities of Conditionals and Conditional Probabilities" (1976b) and its sequel (1986d), and "Desire as Belief" (1988b) and its sequel (1996a). The papers have more in common than merely having a common naming convention. (They're not even Lewis's only sequels; "Lucas Against Mechanism" (1969b) also has a sequel (1979c).) In both cases Lewis aims to defend orthodox Bayesian epistemology against some challenges. And in both cases the argument turns on principles to do with updating. Lewis was throughout his career a Bayesian; he frequently said that the ideal epistemic agent was a Bayesian conditionaliser and utility maximiser. And he defended this position with some gusto.

The conditionals papers concern a position that was gaining popularity before Lewis showed it was untenable. The position in question starts with the idea that a speaker can properly say *Probably, if p, q* iff their subjective probability of *q* given *p* is high. And the position then offers an explanation of this purported fact. The English word 'if' is a binary connective which forms a sentence to be written as $p \rightarrow q$, and it is true in virtue of the meaning of this connective that $Pr(q \mid p) = \Pr(p \rightarrow q)$. So, assuming 'probably' means something like subjective probabilility *Probably, if p, q* means that the subjective probability of $p \rightarrow q$, and, assuming the agent is coherent, that is true just in case the subjective probability of *q* given *p* is high.

Lewis doubted several aspects of this story. He briefly notes in "Adverbs of Quantification" that he didn't think the 'if' in *Probably, if p, q* is a binary connective. But the more telling objection was his proof that there could not be a connective $\rightarrow$ such that for all *p*, *q*, $Pr(q \mid p) = Pr(p \rightarrow q)$. Lewis first argued for this in (1976b), and showed how to weaken some of the assumptions of the argument in (1986d). The effect of Lewis's position was to essentially

end the hope of analysing English 'if' in terms of a binary connective with these probabilistic properties.

The desire papers (1988a; 1996b) are also about the Humean view that motivation requires both a belief and a desire. Lewis aims to attack the anti-Humean position that some beliefs, in particular beliefs that a certain thing is good, can play the functional roles of both beliefs and desires. He argues that this is not, in general, possible. And the argument is that beliefs and desires update in different ways. Or, at least, that anyone who updates their beliefs by conditionalisation and updates their valuation functions in a plausible way, will not be able to preserve any correlation between desire for a proposition being true and belief in that proposition's goodness.

Both of these papers rely on the idea that conditionalisation is a good way to update beliefs. Neither, by the way, rely on the idea that conditionalisation is the only rational way to update beliefs; the arguments go through given merely the permissibility of conditionalising. Many Bayesians hold something stronger, namely that conditionalisation is *the* way to update beliefs. One widely used argument in favour of this position is a so-called 'Dutch Book' argument. This argument shows that if you plan to follow any strategy for revising beliefs other than conditionalisation, and you do follow that strategy, then someone who knows the strategy that you're going to follow can produce a series of bets that will seem favourable to you when each is offered, but which will collectively lead to a sure loss. If you conditionalise, however, no such series of bets can be produced. This argument was introduced to the literature by Paul Teller (1973), who credited it to Lewis. Lewis's own version of the argument did not appear until 1999, in *Papers in Metaphysics and Epistemology*, under the title "Why Conditionalize?" (1999b). This was something he had written as a course handout in 1972, and which had been very widely circulated, and, via Teller's paper, very influential on the development of Bayesian epistemology.

Lewis was an early proponent of one of the two major views about the Sleeping Beauty puzzle. (There is a good description of the puzzle in section 6.3 of the entry on epistemic paradoxes, so I won't repeat the description here.) The puzzle was introduced to the philosophical community by Adam Elga (2000b), who argued that when Beauty woke up, her credence in Heads should be $\frac{1}{3}$. Lewis argued that the correct answer was $\frac{1}{2}$. The core of his argument was that before Beauty went to sleep, her credence in Heads should be $\frac{1}{2}$. That was agreed on all sides. Moreover, nothing happened that surprised Beauty. Indeed, everything happened exactly as she expected it would. Lewis argued that "Only new relevant evidence, centred or uncentred, produces a change in credence" (2001b, 174), and that Beauty got no new evidence. This idea has featured heavily in subsequent work defending the $\frac{1}{2}$ answer to the Sleeping Beauty puzzle.

The Sleeping Beauty puzzle is important for another reason. As the quote above indicates, the puzzle is usually set up in terms of sets of centered worlds, following the work of Lewis we described in section 4.5. The work generated by the puzzle has been one of the reasons that that work, in particular (1979a), has received a large amount of attention in recent years.

7.4 Philosophy of Religion

In "Anselm and Actuality" (1970a), Lewis tries to give as good a formulation of the ontological argument as can be made in modal realist terms. This is a good framework for discussing the ontological argument, since on one interpretation, the argument rests crucially on cross-world comparisons of greatness and the modal realist can make sense of that kind of talk better than

views that reject possible objects. Lewis argues that the principle "A being than which nothing greater can be conceived is possible" is crucially ambiguous. One kind of reading is that the imagined being's greatness in its world is greater than the greatness of any other being in that being's world. That may be true, but it doesn't imply that the being actually exists. Another kind of reading focusses on the imagined being's greatness in this world. It says that there (actually) is a being whose actual greatness is greater than the greatness of any possible being. That entails the conclusion, but is not plausibly true. The broader conclusion here, that the ontological argument derives its persuasive force from an equivocation, is one that has been widely adopted since Lewis's paper.

In "Evil for Freedom's Sake" (1993a), Lewis reflects at length on the free will defence to the problem of evil. Lewis argues that for the defence to work, God must make quite different trade-offs between freedom and welfare than we are usually disposed to make, and our understanding of what freedom consists in, and what divine foreknowledge consists in, must be different to what they currently are.

In "Do We Believe in Penal Substitution?" (1997a), Lewis notes that we only sometimes accept that one person can be properly punished for another's misdeeds. He uses this to raise an interesting difficulty for the Christian idea that Christ died for our sins, suggesting this may not be a form of penal substitution that is normally acceptable.

In "Divine Evil" (2007), Lewis suggests that proponents of the problem of evil should not focus on what God fails to prevent, but on what God does. In orthodox forms of theism, particularly Christianity and Islam, God is presented as perpetrating great evil against sinners of various stripes in the form of extreme punishments in the afterlife. Lewis suggests that a God that does would be so evil that we should not only reject Him, but we may regard those who endorse the divine punishments as themselves somewhat culpable for divine evil. (The published version of this paper was composed by Phillip Kitcher after Lewis's death from notes Lewis made, and conversations Kitcher had with Lewis.)

7.5 Ethics

Lewis is obviously not as well known for his work in ethics as for his work in other areas of philosophy. It was something of a surprise when one of the volumes of his collected papers was called *Papers in Ethics and Social Philosophy* (2000). On the other hand, the existence of this volume indicates that there is a large body of work that Lewis put together in moral philosophy, very broadly construed. The best guide to this work is chapter 8 of Nolan (2005), and I'll follow Nolan very closely here.

As Nolan suggests, the least inaccurate summary of Lewis's ethical positions is that he was a virtue ethicist. Indeed, a focus on virtue, as opposed to consequences, plays a role in his defence of modal realism, as we saw in section 6.4. Nolan also notes that this position is somewhat surprising. Most philosophers who accept views related to Lewis's about psychology and decision-making (in particular, who accept a Humean story about beliefs and desires being the basis for motivation, and who accept some or other version of expected utility maximisation as the basis for rational decision) have broadly consequentialist positions. But not Lewis.

Lewis was also a value pluralist (1984a; 1989b; 1993a). Indeed, this was part of his objection to consequentialism. He rejected the idea that there was one summary judgment we could make about the moral value of a person. In "Reply to McMichael" (1978a) he complains about the utilitarian assumption that "any sort or amount of evil can be neutralized, as if it had never

been, by enough countervailing good —and that the balancing evil and good may be entirely unrelated" (1978a, 85).

In meta-ethics, Lewis defended a variety of subjectivism (1989b). Like many subjectivists, Lewis held that something is valuable for us iff we would value it under ideal circumstances. And he held, following Frankfurt (1971), that valuing something is simply desiring to desire it. What is distinctive about Lewis's position is his view about what ideal circumstances are. He thinks they are circumstances of "full imaginative acquaintance". This has some interesting consequences. In particular, it allows Lewis to say that different goods have different conditions of full imaginative acquaintance. It might, he suggests, be impossible to properly imagine instantiating several different values at once. And that in turn lets him argue that his value pluralism is consistent with this kind of subjectivism, in a way that it might not be consistent with other varieties of subjectivism.

Lewis also wrote several more papers in applied ethics. In two interesting papers on tolerance (1989a; 1989c), he suggests that one reason for being tolerant, and especially of being tolerant of speech we disapprove of, comes from game-theoretic considerations. In particular, he thinks our motivation for tolerance comes from forming a 'tacit treaty' with those with differing views. If we agree not to press our numerical superiority to repress them when we are in the majority, they will do the same. So tolerating opposing views may be an optimal strategy for anyone who isn't sure that they will be in the majority indefinitely. In these works it is easy to see the legacies of Lewis's early work on philosophical lessons to be drawn from game theory, and especially from the work of Thomas Schelling.

7.6 Applied Metaphysics

There's much more that could be said about Lewis's contributions to philosophy, but we'll end with a discussion of two wonderful pieces of applied metaphysics.

In "The Paradoxes of Time Travel" (1976a), Lewis discusses the many complicated philosophical issues about time travel. He discusses temporal parts, personal identity, causation and causal loops, free will, and the complications arising from our many different modal concepts. In some cases he uses the canvas provided to illustrate his own take on the metaphysical issues that arise. But in some cases he notes that the problems that arise are problems for everyone.

"Holes" (Lewis and Lewis, 1970) was co-written with Stephanie Lewis. In it they discuss, in dialog form, some of the metaphysical issues that holes generate. One of the characters, Argle, wants to eliminate holes from his ontology, and the paper goes over what costs must be met to make this form of nominalism work. The other character, Bargle, pushes Argle to clarify his commitments, and in doing so draws out many details of the nominalist framework. The case is of some interest in itself, but it is also, as the authors note at the end, a useful case-study in the kind of moves nominalists can make in eliminating unwanted ontology, and the costs of those moves.

Each paper can be, and indeed often has been, used for introducing complicated metaphysical issues to students. The papers are, like many of Lewis's papers, widely anthologised. They are both excellent illustrations of the fact that, as well as being a wonderful philosopher, Lewis was one of the best philosophical writers of his time.

Humean Supervenience

1 What is Humean Supervenience?

As with many aspects of David Lewis's work, it is hard to provide a better summary of his views than he provided himself. So the following introduction to what the Humean Supervenience view is will follow the opening pages of Lewis (1994a) extremely closely. But for those readers who haven't read that paper, here's the nickel version.

Humean Supervenience is the conjunction of three theses.

1. **Truth supervenes on being** (Bigelow, 1988). That is, all the facts about a world supervene on facts about which individuals instantiate which fundamental properties and relations.
2. **Anti-haeccaetism**. All the facts about a world supervene on the distribution of qualitative properties and relations; rearranging which properties hang on which 'hooks' doesn't change any facts.
3. **Spatio-temporalism**. The only fundamental relations that are actually instantiated are spatio-temporal, and all fundamental properties are properties of points or point-sized occupants of points.

The first clause is a core part of Lewis's metaphysics. It is part of what it is for some properties and relations to be fundamental that they characterize the world. Indeed, Lewis thinks something stronger, namely that the fundamental properties and relations characterize the world *without redundancy* (Lewis, 1986b, 60). This probably isn't true, for a reason noted in Sider (1993). Consider the relations *earlier than* and *later than*. If these are both fundamental, then there is some redundancy in the characterisation of the world in terms of fundamental properties and relations. But there is no reason to believe that one is fundamental and the other isn't. And it is hard to see how we could give a complete characterisation of the world without either of these relations. So we'll drop the claim that the fundamental properties relations characterise the world without redundancy, and stick to the weaker claim, namely that the fundamental properties and relations characterize the world completely.

The second clause is related to Lewis's counterpart theory. Consider what it would be like for anti-haeccaetism to fail. There would have to be two worlds, with the same distribution of qualitative properties, but with different facts obtaining in each. These facts would have to be non-qualitative facts, presumably facts about which individual plays which role. So perhaps, to use a well-known example, there could be a world in which everything is qualitatively as it is in this world, but in which Barack Obama plays the Julius Caeser role, and vice versa. So Obama conquers Gaul and crosses the Rubicon, Caeser is born in Hawai'i and becomes President of the United States. But what could make it the case that the Gaul-conqueror in that world is really *Obama's* counterpart, and not Caeser's? Nothing qualitative, and nothing else it seems is available. So this pseudo-possibility is not really a possibility. And so on for all other counterexamples to anti-haeccaetism.

[†] In progress. Commissioned for a handbook on David Lewis.

The third clause is the most striking. It says there are no fundamental relations beyond the spatio-temporal, or fundamental properties of extended objects. If we assume that 'properties' of objects with parts are really relations between the parts, and anything extended has proper parts, then the second clause reduces to the first. I think it isn't unfair to read Lewis as holding both those theses.

Since for Lewis the fundamental qualities are all intrinsic, the upshot is that the world is characterized by a spatio-temporal distribution of intrinsic qualities. As Lewis acknowledged, this was considerably more plausible given older views about the nature of physics than it is now. We'll return to this point at great length below. But for now the key point to see the kind of picture Humean Supervenience offers. The world is like a giant video monitor. The facts about a monitor's appearance supervene, plausibly, on intrinsic qualities of the pixels, plus facts about the spatial arrangement of the pixels. The world is 4-dimensional, not 2-dimensional like the monitor, but the underlying picture is the same.

2 Supervenience

Given the name Humean Supervenience you might expect it to be possible to state Humean Supervenience as a supervenience thesis. But this turns out to be hard to do. Here is one attempt at stating Humean Supervenience as a supervenience thesis that is happily clear, and unhappily false.

Strong Modal Humean Supervenience For any two worlds where the spatio-temporal distribution of fundamental qualities is the same, the contingent facts are the same.

But Humean Supervenience does not make a claim this strong. It is consistent with Humean Supervenience that there could be fundamental non-spatio-temporal relations. The only thing Humean Supervenience claims is that no such relations are instantiated. In a pair of possible worlds where there are such relations, and the relations vary but the arrangement of qualities is the same, **Strong Modal Humean Supervenience** will fail. In the Introduction to Lewis (1986c), he suggested the following weaker version.

Local Modal Humean Supervenience For any two worlds at which no alien properties or relations are instantiated, if the spatio-temporal distribution of fundamental qualities is the same at each world, the contingent facts are also the same.

An alien property(/relation) is a fundamental property(/relation) that is not actually instantiated. So this version of Humean Supervenience says that to get a difference between two worlds, you have to either have a change in the spatio-temporal arrangement of qualities, or the instantiation of actually uninstantiated fundamental properties or relations.

But Lewis eventually decided that wouldn't do either. In response to Haslanger (1994), he conceded that enduring objects would generate counterexamples to **Local Modal Humean Supervenience** even if there were no alien properties or relations. So he fell back to the following, somewhat vaguely stated, thesis. (See Lewis (1994a) for the concession, and **?** for an argument that he should not have conceded this to Haslanger.)

Familiar Modal Humean Supervenience In any two "worlds like ours" (Lewis, 1994a, 475), if the spatio-temporal distribution of fundamental qualities is the same at each world, the contingent facts are also the same.

What's a "world like ours'? It isn't, I fear, entirely clear. But this doesn't matter for the precise statement of Humean Supervenience. The three theses in section 1 are clear enough, and state what Humean Supervenience is. The only difficulty is in stating it as a *supervenience* thesis.

3 What is Perfect Naturalness?

That definintion does, however, require that we understand what it is for some properties and relations to be *fundamental*, or, as Lewis put it following his discussion in Lewis (1983b), *perfectly natural*. The perfectly natural properties and relations play a number of interconnected roles in Lewis's metaphysics and his broader philosophy.

Most generally, they characterise the difference between real change and 'Cambridge change', and the related difference between real similarity, and mere sharing of grue-like attributes. This somewhat loose idea is turned, in *Plurality*, into a definition of duplication.

> …two things are duplicates iff (1) they have exactly the same perfectly natural properties, and (2) their parts can be put into correspondence in such a way that corresponding parts have exactly the same perfectly natural properties, and stand in the same perfectly natural relations. (Lewis, 1986b, 61)

The intrinsic properties are then defined as those that are shared between any two (possible) duplicates. So, as noted above, Humean Supervenience says that the spatio-temporal distribution of intrinsic features of points characterises worlds like ours.

I've gone back and forth between describing these properties as fundamental and describing them as perfectly natural. And that's because for Lewis, the perfectly natural properties are in a key sense fundamental. For reasons to do with the nature of vectorial properties, I think this is probably wrong (Weatherson, 2006). That is, we need to hold that some derivative properties are perfectly natural in order to get the definition of intrinsicness terms of perfect naturalness to work. But for Lewis, the perfectly natural properties and relations are all fundamental.

Part of what Lewis means by saying that some properties are fundamental is that all the facts about the world supervene on the distribution. (This is Bigelow's thesis that truth supervenes on being.) But I think he also means something stronger. The non-fundamental facts don't merely supervene on the fundamental facts; those non-fundamental facts are true *because* the fundamental facts are true, and *in virtue of* the truth of the fundamental facts.

The perfectly natural properties play many other roles in Lewis's philosophy besides these two. They play a key role in the theory of laws, for instance. They are a key part of Lewis's solution to the New Riddle of Induction (Goodman, 1955). And they play an important role in Lewis's theory of content, though just exactly what that role is is a matter of some dispute. (See Sider (2001a) and Weatherson (2003b) for one interpretation, and Schwarz (2009) for a conflicting interpretation.)

Now it is a pretty open question whether any one division of properties can do all these roles. One way to solve the New Riddle (arguably Lewis's way, though this is a delicate question of interpretation) is to be a dogmatist (in the sense of Pryor (2000)) about inductive projections

involving a privileged class of properties. Lewis's discussion of the New Riddle at the end of Lewis (1983b) sounds like he endorses this view, with the privileged class being the very same class as fundamentally determines the structure of the world, and makes for objective similarity and difference. But why should these classes be the same? It might make more sense to, for instance, endorse dogmatism about inductive projections of *observational* properties, rather than about microphysical properties.

Lewis doesn't attempt to give a theoretically neutral definition of the perfectly natural properties. Rather, the notion of a perfectly natural property is introduced by the theoretical role it serves. But that theoretical role is very ambitious, covering many areas in metaphysics, epistemology and the theory of content. We might wonder whether claims like Humean Supervenience have any content if it turns out nothing quite plays that theoretical role. I think there is still a clear thesis we can extract, relying on the connection between intrinsicness and naturalness. It consists of the following claims:

- There is a small class of properties and relations such that the contingent facts at any world supervene on the distribution of these properties and relations.
- Each of these properties is an intrinsic property.
- At the actual world, the only relations among these which are instantiated are spatiotemporal, and all the contingent facts supervene not merely on the distribution of fundamental qualities and relations, but also on the distribution of fundamental qualities and relations *over points and point-sized occupants of points*.

Those theses are distinctively Lewisian, they are clearly entailed by Humean Supervenience as Lewis's conceives of them, they are opposed in one way or another by those who take themselves to reject Humean Supervenience, but they are free of any commitment to there being a single class of properties and relations that plays all the roles Lewis wants the perfectly natural properties and relations to play. So from now on, when I discuss the viability of Humean Supervenience, I'll be discussing the viability of this package of views.

4 Humean Supervenience and other Humean Theses

Lewis endorsed many views that we might broadly describe as 'Humean'. Of particular interest here are the following three.

- **Humean Supervenience**.
- **Nomological Reductionism**. Nomological properties and relations (including lawhood, chance and causation) are not among the fundamental properties and relations.
- **Modal Combinatorialism**. Roughly, anything can co-exist with anything else.

We've stated Modal Combinatorialism extremely roughly, and will persist with using a fairly informal version of it throughout. For an excellent study of more careful versions of it, see Nolan (1996). But those details aren't as important to this debate. What is important for now is that all three of these theses are associated with what are known as Humean approaches to metaphysics in the contemporary literature. But how closely connected are they to each other, or for that matter to Hume.

One question about Humean Supervenience is just how it connects to the work of the historical Hume. This would be a little easier to answer if there was a broad scholarly consensus that Hume actually believed the kind of simple regularity thesis of causation that Lewis attributes to him at the start of Lewis (1973a). But it isn't clear that this is Hume's view (Strawson, 2000). What is true is that Hume was sceptical that we could know more about causation than that it was manifested in certain distinctive kinds of correlations. But it is a further step to say that Hume inferred that causation just consists of these distinctive kinds of correlations.

A second question is how Humean Supervenience, which perhaps should be referred to as so-called "Humean Supervenience", or perhaps even better as "Lewisian Supervenience", relates to the kind of regularity theory that Lewis attributes to Hume, or to the prohibition on necessary connections between distinct existences that underlies Modal Combinatorialism. Lewis seemed to see the three theses as related. Here he is explaining how he chose to name Humean Supervenience (and recall that this *isn't* backed up by any detailed exegesis of Hume).

> Humean Supervenience is named in honour of the great denier of necessary connections. It is the doctrine that all there is to the world is a vast mosaic of local matters of particular fact just one little thing and then another. (Lewis, 1986c, ix)

This is a slightly confusing passage, since it isn't clear why a violation of Humean Supervenience would constitute a necessary connection of any kind. We will return to this point below. But it does seem to make clear that Lewis thought that Humean Supervenience and Modal Combinatorialism were connected, since Modal Combinatorialism is much more closely connected to the denial that they can be necessary connections between distinct existences.

Compare how Lewis introduces Humean Supervenience when discussing the role of possible worlds in formulating trans-world supervenience theses in *Plurality*.

> Are the laws, chances, and causal relationships nothing but patterns which supervene on this point-by-point distribution of properties? Could two worlds differ in their wars without differing, somehow, somewhere, in local qualitative character? (I discuss this question of 'Humean Supervenience', inconclusively, in the Introduction to my *Philosophical Papers*, volume II.) (Lewis, 1986b, 14)

This seems to connect Humean Supervenience closely to Nomological Reductionism, since it makes the reducibility of the nomological properties and relations central to the question of whether Humean Supervenience is true. We can also, I think, see Lewis connecting Modal Combinatorialism and Nomological Reductionism in a later passage in *Plurality* where he discusses why he doesn't believe that laws are necessary truths.

> Another use of [Modal Combinatorialism] is to settle – or as opponents might say, to beg – the question whether the laws of nature are strictly necessary. They are not … Episodes of bread-eating are possible because actual; as are episodes of starvation. Juxtaposed duplicates of the two, on the grounds that anything can follow anything; here is a possible world to violate the law bread nourishes. … It is no surprise that [Modal Combinatorialism] prohibited strictly necessary connections between distinct existences. What I have done is to take a Humean

> view about laws and causation, and use it instead as a thesis about possibility. Same thesis, different emphasis. (Lewis, 1986b, 91)

So for Lewis, these three theses are meant to be closely connected. And it is true that in the contemporary literature all three of them are frequently described as 'Humean' theses. (Or at least they are so described in metaphysics and philosophy of science; again, we're bracketing questions of historical interpretation here.) But on second glance, it isn't as clear what the connection between the three theses could amount to. One immediate puzzle is that Humean Supervenience is for Lewis a *contingent* thesis, while the other two theses are necessary truths. The accounts of causation, lawhood and chance that he gives in defending Nomological Reductionism are clearly meant to hold in all kinds of worlds, not just worlds like ours. (Consider the amount of effort that is spent in Lewis (2004a) at defending the theory of causation from examples involving wizards, action at a distance and so on.) And the formulation of Modal Combinatorialism in *Plurality* leaves little doubt that it is meant to be necessarily true.

This difference in modal status means that the theses can't be in any way equivalent. But you might think that they are in some way reinforcing. Even that isn't so clear. Consider the most dedicated kind of denier of Modal Combinatorialism, namely the fatalist who thinks that every truth is a necessary truth. She will *endorse* Humean Supervenience. After all, she thinks that all the truths about the world supervene on any category of truths whatsoever, so they'll supervene on intrinsic properties of point-sized objects.

In the other direction, failures of Humean Supervenience don't motivate compromising Modal Combinatorialism. Imagine a world where occasionally there are pairs of people who can know what each other is thinking, even though there is no independent informational chain between the two of them. It is just that a telepathic connection exists. Moreover, there is no rhyme or reason to when a pair of people will be telepathic; it is simply the case that some pairs of people are. In such a world, it is plausible that *being a telepathic pair* will be a fundamental relation. That's not a problem for Humean Supervenience, since there aren't any such pairs in this world. But it does mean Humean Supervenience is false in that world.

Assume that Daniels and O'Leary are a telepathic pair. Any duplication of the pair of them will also be telepathic, since by Lewis's preferred definition of duplication, duplication preserves all fundamental properties and relations. Does that mean there's a necessary connection between Daniels and O'Leary? Not really. The spirit of Modal Combinatorialism is that you can duplicate any parts of any worlds, and combine them. One part of our world is Daniels. A duplicate of him need not include any telepathic connection to O'Leary; indeed, he has duplicates in worlds in which O'Leary is absent. Another part of the world is O'Leary; duplicates of him need not include a connection to Daniels. Putting the two together, there is a world where there are duplicates of Daniels and O'Leary, but no telepathic connection between the two. So Modal Combinatorialism suggests that even when Humean Supervenience fails, there won't be a necessary connection between distinct objects. So Humean Supervenience really isn't that important to the idea that there are no necessary connection between distinct existences.

What's closer to the truth, I think, is that Humean Supervenience is *interesting* because of Modal Combinatorialism. If Modal Combinatorialism fails, then Humean Supervenience doesn't capture anything important. In particular, it doesn't capture the idea that the nomic is somehow less fundamental than (some features of) the non-nomic. It is only given Modal

Combinatorialism that we can make these kinds of priority claims in modal terms. Think about the philosopher who denies Modal Combinatorialism on the grounds that laws of nature are necessarily true. That philosopher will say that the laws supervene on the distribution of intrinsic properties of points, because the laws supervene on any set of facts that you like. But they will deny that this makes the distribution of intrinsic properties of points more fundamental than the laws. It is only given Modal Combinatorialism that we can claim that supervenience theses are any guide whatsoever to fundamentality.

What about the connection between Nomological Reductionism and Humean Supervenience? It can't be equivalence, since Lewis agrees that Humean Supervenience fails in worlds in which Nomological Reductionism is true. For the same reason, it can't be that failures of Humean Supervenience entail failures of Nomological Reductionism. What about the other direction? Could we imagine Nomological Reductionism failing while Humean Supervenience holds? I think this is a coherent possibility, but not at all an attractive one. (Compare, in this respect, the discussion of theories that "qualify technically as Humean" at (Lewis, 1994a, 485).) It requires that some of the irreducible, nomological properties be intrinsic properties of point-sized objects. Well, we could imagine two worlds where F and G are co-extensive, intrinsic properties of points, and in one of them it is a law that all Fs are Gs, and in the other it is a law that all Gs are Fs, and there are further intrinsic properties of all the points which are F and G which underlie these laws without making a difference to any of the other facts. So we imagine that the property *being F in virtue of being G* is held by all these things in one world but not in the other, and this is a fundamental perfectly natural property. I don't think any of this is literally inconsistent, and I think filling out the details could give us a way for Nomological Reductionism to fail while the letter of Humean Supervenience holds. But it would clearly violate the spirit of Humean Supervenience and it isn't clear why we should believe in such 'possibilities' anyway.

So in practice, I think that any philosopher who rejects Nomological Reductionism is probably going to want to reject Humean Supervenience. And I think that Lewis saw some of the deepest challenges to Humean Supervenience as coming from threats to Nomological Reductionism. In particular, Lewis thought that the biggest challenges to Humean Supervenience came from the difficulties in providing a reductive account of chance, and the appeal of non-reductive series of causation.

The difficulties in providing a reductive account of chance are discussed at length in the introduction to Lewis (1986c), and in the only paper that has 'Humean Supervenience' in its title, i.e., Lewis (1994a). Here is a quick version of the problem. Chances are not fundamental, so they must supervene on the distribution of qualities. At least in the very early stages of the universe, there aren't enough facts about the distribution of qualities in the past and present to form a suitable subvenient base for the chances. So whether the chance of p is x or y will, at least some of the time, depend on how the future of the world turns out. Now let p the proposition that tells the full story about the future of the world. And assume that p is a proposition such that what its chance is depends on how that future goes. If it goes the way p says it will go, the chance of p is x; if it goes some other way, the chance of p is y. Given a Humean theory of chance, Lewis says that this is going to be possible.

But now there's a problem. What Lewis calls the Principal Principle says that if we know the chance of p is y, and have no further information, then our credence in p should be y. But in this case, if we knew the chance of p was y, we could be sure that p would not

obtain. So our credence in p should be 0. Here we seem to have reached a contradiction, and it is a contradiction to Lewis for a long time feared undermined the prospect of giving a reductive account of chance. The solution he eventually settled on in Lewis (1994a) was to slightly modify the Principal Principle, with the modification being designed to make very little difference in regular cases, but avoid this contradiction.

Lewis discusses the appeal of non-reductive theories of causation in several places, most notably for our purposes Lewis (2004a) and Lewis (2004c). Much of his attention is focused on the theory developed by Peter Menzies (1996). Menzies suggests that causation is the intrinsic relation that does the best job of satisfying folk platitudes about causation. A consequence of Menzies's view is that there is something that makes a difference to the intrinsic properties of pairs of causes and effects which doesn't supervene on either the intrinsic properties of the two ends of the causal chain, or on the spatio-temporal relations that hold between them. This something will either be causation or will be something on which causation depends. Either way there is a problem for Humean Supervenience, since there will have to be a perfectly natural relation that is not spatio-temporal.

Lewis's response is to raise problems for the idea that causation could be an intrinsic relation. One class of worries concerns the very idea that causation could be a relation. Lewis says that absences can be causes and effects, but absences can't stand in any relations, so causation must not be a relation. Another class of worries concerns the idea that causation could be intrinsic. Causation by double prevention, says Lewis, doesn't look like it could be intrinsic. But intuitively there could be causation by double prevention. Yet another class of worries concerns the idea that causation could be a natural relation, or that there could be any one thing that satisfies all the platitudes about causation. The vast array of different ways in which causes can bring about their effects in the actual world, he says, undermines this possibility.

Note that in both cases Lewis defends Humean Supervenience simply by defending Nomological Reductionism. So I think it is fair to say that there's a close connection between the two in Lewis's overall theory.

5 Why Care about Humean Supervenience

As is well-known, some surprising results in quantum mechanics suggest that entanglement relations are somehow fundamental (Maudlin, 1994). This suggests that Humean Supervenience is actually false. If that's right, why should we care about philosophical arguments for Humean Supervenience? Lewis's response to this challenge is somewhat disconcerting.

> Really, what I uphold is not so much the truth of Humean Supervenience as the *tenability* of it. If physics itself were to teach me that it is false, I wouldn't grieve.
>
> That might happen: maybe the lesson of Bell's Theorem is exactly that … But I am not ready to take lessons in ontology from quantum physics as it now is. … If, after [quantum theory has been cleaned up], it still teaches non-locality, I shall submit willingly to the best of authority.
>
> What I want to fight are *philosophical* arguments against Humean Supervenience. When philosophers claim that one or another commonplace feature of the world cannot supervene on the arrangement of qualities, I make it my business to resist. Being a commonsensical fellow (except where unactualised possible worlds

> are concerned) I will seldom deny that the features in question exist. I grant their existence, and do my best to show how they can, after all, supervene on the arrangement of qualities. (Lewis, 1986c, xi)

We can, I think, dismiss the point about quantum physics as it was in 1986. The theory has been cleaned up in just the way Lewis wanted, and the claims about non-locality remain. Indeed, by the end of his life Lewis was willing to take lessons in ontology from quantum physics. See, for example, Lewis (2004b). So what is at issue here is whether or not there are philosophical arguments against Humean Supervenience.

But at this point we might wonder why we should care. If a theory is false, what does it matter whether its falsehood is shown by philosophy or by physics? We might compare the dismissive attitude Lewis takes towards Plantinga's attempts to show that reconstructions of the problem of evil as an argument do not rely solely on things provable in first-order logic (Lewis, 1993a).

The answer I offered in Weatherson (2009) was that the philosophical defence of Humean Supervenience was connected to the point of the last paragraph quoted above. Lewis wanted to save various features of our commonsensical picture of the world. And he wanted to do this without saying that philosophical reflection showed us that the picture of the world given to us by signs of somehow incomplete. He wanted to defend what I called 'compatibilism', something that I contrasted with eliminativism and expansionism. The eliminativists want to say that science shows us that some commonsensical feature of reality doesn't really exist. (See, for example, Churchland (1981) for eliminativism about folk psychological states.) The expansionists want to say that since science (or at least physics) doesn't recognise certain features of reality, but they obviously exist, we need to posit that science (or at least physics) is incomplete. There are many stripes of philosophical expansionists, from theists to dualists to believers in agent causation.

Lewis wasn't averse in principle to either eliminativism or expansionism. One could, depending on exactly how one interpreted folk theory and science, classify him as an eliminativist about gods, and an expansionist about unactualised possible worlds. But his first tendency was always to support compatibilism. Compatibilists face what Frank Jackson (1998) called the 'location problem'. They have to show where the commonsensical features are located in the scientific picture. That is, they have to show how to reduce (in at least some sense of 'reduce') or commonsensical concepts to scientific concepts. (Many compatibilists may bristle at the idea that they have to be reductionists; in recent decades the world has abounded with 'non-reductive physicalists', who are precisely compatibilists in my sense, but who reject what they call 'reductionism'. But as Lewis (1994b) argued, these rejections often turn on reading too much into the notion of reduction. For that reason, Lewis would not have objected to being described as a reductionist about many everyday concepts.)

One way to perform such a reduction would be to wait until the best scientific theory is developed, and show where within it we find minds, meanings, morals and all the other exciting features of our ordinary worldview. But that could take a while, and philosophers could use something to do while waiting. In the meantime we could look for a recipe that should work no matter what physical theory the scientists settle on, or at least should work in a very wide range of cases. I think we can see Lewis's defence of Humean Supervenience as providing such a recipe.

It is important to note here that Lewis's defence of Humean Supervenience was largely *constructive*. He didn't try to give a proof that there couldn't be more to the world than the arrangement of local qualities. At least, he didn't rest a huge amount of weight on such arguments. The arguments we will look at below for a functional construal of the nomological are, perhaps, hints at arguments of this type. But, in general, Lewis defended Humean Supervenience by explicitly showing where the ordinary concepts fitted in to a sparse physical picture of reality, under the assumption that physics tells us that the world consists of nothing but a spatio-temporal arrangement of intrinsic qualities.

Now physics tells us no such thing. But it shouldn't matter. If the recipe Lewis provides works in the case of the 'Humean' world, it should also work in the world physics tells us we actually live in. The reduction of laws to facts about the distribution of fundamental qualities, and the reduction of chances and counterfactual dependencies to facts about laws, and the reduction of causation to facts about chances and counterfactual dependencies, and the reduction of mind to facts about causation and the distribution of qualities, and the reduction of value to facts about minds, and so on are all independent of whether physics tells us that we have to recognise relationships like entanglement as fundamental. In other words, if we can solve the location problem for the Humean world, we can solve it for the actual world. And solving the location problem is crucial to defending compatibilism. And whether it is possible to defend compatibilism is a central concern of metaphysics.

I quoted above a passage from 1986 in which Lewis links Humean Supervenience to compatibilism. It's worth noting that he returns to the point in 1994.

> The point of defending Humean Supervenience is not to support reactionary physics, but rather to resist philosophical arguments that there are more things in heaven and earth in physics has dreamt of. Therefore if I defend the *philosophical* tenability of Humean Supervenience, that defence can doubtless be adapted to whatever better supervenience thesis may emerge from better physics. (Lewis, 1994a, 474)

That is, the defence of Humean Supervenience just is part of the argument against expansionism, and hence for compatibilism. That was the defence I offered in Weatherson (2009) for the interest of Lewis's defence of Humean Supervenience, even if it were to turn out that Humean Supervenience was refuted by physics. I still think much of it is correct. In particular, I still think that Lewis wanted to defend compatibilism, and that the defence of Humean Supervenience is key to the defence of Humean Supervenience. Indeed, I think there is pretty strong textual evidence that it was a major part of Lewis's motivation for defending Humean Supervenience. But this explanation of why the defence of Humean Supervenience is significant can't explain why Lewis was so worried about the failures of Humean theories of chance. After all, if all we are trying to do is show that science and commonsense are compatible, we could just take chances to be one of the fundamental features of reality given to us by science. There isn't any need, from the perspective of trying to reconcile science and common sense, to give a reductive account of chance. Yet Lewis clearly thought that giving a reductive account of chance was crucial to the defence of Humean Supervenience. As he said,

> There is one big bad bug: chance. It is here, and here alone, that I fear defeat. But if I'm beaten here, then the entire campaign goes kaput. (Lewis, 1986c, xiv)

I now think that attitude is very hard to explain if my earlier views about the significance of Humean Supervenience are entirely correct. The natural conclusion is that there is something *more* that the defence of Humean Supervenience is supposed to accomplish. One plausible interpretation is that what it is supposed to accomplish is a vindication of the idea that the key nomological concepts are, in a sense, descriptive. It's easiest to say what this sense is by contrasting it with the kind of view that Lewis rejected.

We're all familiar with the standard story about 'water'. Our ordinary usage of the term latches onto some stuff in the physical world. That stuff is H_2O. Some people think that's because our ordinary usage determines a property which H_2O satisfies, others because we demonstratively pick out H_2O in ordinary demonstrations of what it is we're talking about when we use the term 'water'. Either way, we get to be talking about H_2O when we use the word 'water', even if we are so ignorant of chemistry that we can't tell hydrogen and oxygen apart. Moreover, our term continues to pick out 'water' even in worlds that are completely free of hydrogen and oxygen, and even if such worlds have other stuff that plays a very similar functional role to the role water plays in the actual world.

Lewis was somewhat sceptical of this standard story about 'water' (Lewis, 2002b). He thought that the ordinary term was ambiguous between our usage on which it picked out H_2O, and usage on which it picked out a role, a role that happens to be played by H_2O in the actual world but which could be played by other substances in other worlds. But if he thought the standard story about 'water' was at best, part right, he thought applying a similar story to 'law', 'cause' and 'chance' was wildly implausible.

If such a story were right, then we would expect to find worlds where there was some relation other than causation which played the causal role. Since the actual world is physical, any world in which nonphysical things stand in the kind of relations that causes and effects typically stand in should do. So, for instance, if we have a world where the castings of spells are frequently followed by transformations from human to toad form, we should have a world where spells don't cause such transformations but rather the spellcasting and the transformation stand in a kind of fool's cause relationship. But we see no such thing. In such magical worlds, spells cause transformations.

So whatever causation is, it doesn't look to be the kind of thing whose essence can be discovered by physics. Physics couldn't tell us anything about the essence of the relationship between the spell and the transformation into a toad. But, we think, physics can tell us a lot about the fundamental properties and relations are instantiated in the actual world. So causation must not be one of them.

Lewis has a number of other arguments against anti-descriptivist views about individual nomological concepts. These arguments strike me as rather strong in the case of lawhood and causation, and less strong in the case of chance.

If *being F* and *being F in virtue of a law* are both fundamental properties, then a plausible principle of modal recombination would suggest they could come apart. But they cannot; or at least they cannot in one direction. We want *being F in virtue of a law* to entail *being F*. That's easy if lawhood is defined in terms of fundamental properties of things; but it's hard to see how it could be if lawhood itself is fundamental (Lewis, 1986c, xii).

A similar argument goes for causation. Assume that causation is a fundamental intrinsic relation that holds between things at different times. Consider, for instance, the causal relationship which holds between a throw of a rock (call it t) and the shattering of the window

(call it s). As we noted above in the case of Daniels and O'Leary, several applications of Modal Combinatorialism suggest that there will be a world just like this one in which t is followed by s, but in which t does not cause s. But such a world seems to be impossible. As we also noted above, such a view runs into trouble with causation by double prevention, which does not look to be intrinsic.

The last two paragraphs have been extremely quick arguments, but in both cases it seems to me that they can be tightened up so as to provide good arguments for some kind of descriptivist stance towards laws and causation. Chance is another matter.

The first problem is that recombination arguments if anything point *away* from descriptivism about chance. Any such account will imply that chances can't, in general, point too far away from frequencies. But recombination arguments suggest that chances and frequencies can come arbitrarily far apart. Consider some particular event type e that has a one-half chance of occurring in circumstances c. Start with a world where c occurs frequently, and about half the time it is followed by e. Now use recombination to generate a world where all the $c \wedge \neg e$ events are deleted, so c is always followed by e. Unless we add a lot of bells and whistles to our theory of chance, it will no longer be the case that the chance of e given c is one-half. That is odd; we can't simply take the first circumstance where c occurred and at that moment there was a one-half chance of it being followed by e, and patch it into an arbitrary world. BIGELOW et al. (1993) turn this idea into a more careful argument against descriptivism about chance. They say that chances should satisfy the following principle. (In this principle, Ch is the chance function, and various subscripts relativise it to times and worlds.)

> Suppose $x > 0$ and $Ch_{tw}(A) = x$. Then A is true in at least one of those worlds w' that matches w up to time t and for which $Ch_t(A) = x$. (BIGELOW et al., 1993, 459)

That is, if the chance of A at t is x, and $x > 0$, then A could occur without changing the history prior to t, and without changing the chance of A at t. This seems like a plausible principle of chance, but it entails the not-so-Humean view that chances at t supervene on history to t, not on the full state of the world.

Now as it turns out Lewis *doesn't* rest on recombination arguments against rival views of chance, and in my view he is wise to do so. Instead he rests on epistemological arguments. He takes the following two things to be data points.

1. Something like the Principal Principle is true. The original Principal Principle said that if you knew the chance of p at t was x, and didn't have any 'inadmissible' information (roughly, information about how the world developed after t), then your credence in p should be x. Lewis tinkered with this slightly, as we noted above, but he took it to be a requirement on a theory of chance that the Principal Principle turn out at least roughly right.
2. The correct theory of chance will *explain* the Principal Principle.

Lewis frequently wielded this second requirement against rival theories of chance. Here's one example.

> I can see, dimly, how it might be rational to conform my credences about outcomes to my credencs about history, symmetries and frequencies. I haven't the faintest notion how it might be rational to conform my credences about outcomes to my credences about some mysterious unHumean magnitude. Don't try to take away the mystery my saying that this unHumean magnitude is none other than *chance*! (Lewis, 1986c, xv)

But this also seems like a weak argument. For one thing, chances are actually correlated very well with frequencies, and this correlation does not look at all accidental. It seems very plausible to me that we should line up our credences with things that are actually correlated well with frequencies. But, you might protest, shouldn't we have an explanation of why the Principal Principle is an *a priori* principle of rationality? I think that before we ask for such an explanation, we should check how confident we are that the Principal Principle, or anything else, is part of an *a priori* theory of rationality. I'm not so confident that we'll be able to do this (Weatherson, 2005, 2007).

There are other replies too that we might make. It seems plausible that we should minimise the expected inaccuracy of our credences (Joyce, 1998). This is true when we consider not just the *subjective* expected inaccuracy of our credences, but the *objective* expected inaccuracy of our credences. That is, when we calculate the expected inaccuracy of someone's credences, using chances as the probabilities for generating the expectations, it is good if this expected inaccuracy is as low as possible. But, assuming that we are using a proper scoring rule for measuring the accuracy of credences, this means that we must have credences match chances.

More generally, I'm very sceptical of theories that insist our metaphysics be designed to have complicated epistemological theses fall out as immediate consequences. Rationality requires that we be inductivists. Why is that? Here's a bad way to go about answering it: find a theory of persistence that makes induction obviously rational, and then require our metaphysics to conform to that theory. I don't think you'll get a very good theory of persistence that way, and, relatedly, you won't get a very Lewisian theory of persistence that way. The demand that the theory of chance play a central role in an explanation of the Principal Principle strikes me as equally mistaken.

If what I've been saying so far is correct, then chance interacts with the motivation for Humean Supervenience in very different ways to how laws and causation interact. Neither of the two kinds of motivations for defending Humean Supervenience against philosophical attacks provides us with good reason to leave chances out of the subvenient base on which we say all contingent facts supervene. This is not to yet offer anything like a positive argument for chances to be part of the fundamental furniture of reality. Rather, what I've argued here is that a metaphysics that takes chances as primitives would not be as far removed from a recognisably Lewisian metaphysics as a metaphysics that takes laws or causes as primitive, let alone one that takes mind, meanings or morals as primitive.

6 Points, Vectors and Lewis

The other main point from the discussion of the previous section is that the fact that quantum mechanics raises problems for Humean Supervenience does not undercut the philosophical significance of Lewis's defence of Humean Supervenience. But is Humean Supervenience even compatible with classical physics? Perhaps not.

> Even classical electromagnetism raises a question for Humean Supervenience as I stated it. Denis Robinson (1989) has asked: is a vector field an arrangement of local qualities? I said qualities were intrinsic; that means they can never differ between duplicates; and I would have said offhand that two things can be duplicates even if they point in different directions. May be this last opinion should be reconsidered, so that vector-valued magnitudes may count as intrinsic properties. What else could they be? Any attempt to reconstruct with them as relational properties seems seriously artificial. (Lewis, 1994a, 474)

The opinion that the Lewis proposes to discard here seems more than an offhand judgement. It seems to follow from the very way that we introduce the notion of duplication. Here is Lewis's own attempt to introduce the notion.

> We are familiar with cases of approximate duplication, e.g., when we use copying machines. And we understand that if these machines were more perfect than they are, the copies they made would be perfect duplicates of the original. Copy and original would be alike in size and shape and chemical composition of the ink marks and the paper, alike in temperature and magnetic alignment and electrostatic charge, alike even in the exact arrangement of their electrons and quarks. Such duplicates would be exactly alike we say. They would match perfectly, they would be qualitatively identical, they would be indiscernible. (Lewis, 1983b, 355)

If Lewis is right that vector-valued magnitudes may count as intrinsic properties, then there is yet another condition that the perfect copying machine must satisfy. The original and the duplicate must be parallel. This isn't the case in most actual copying machines. Usually, the original is laid flat, while the duplicate is a small angle to make it easier to collect. This is a feature, not a bug. It is not a way in which the machine falls short of perfect copying. But if vector-valued magnitudes are intrinsic qualities, and duplicates share their intrinsic qualities, it would be. So Lewis is wrong to think that these vector-valued magnitudes may be intrinsic.

Moreover, the little argument that Lewis gives seems to rest on a category mistake. What matters here is the division of properties into intrinsic and extrinsic. But the properties on the kind of things that can be relational or non-relational. As Humberstone (1996) shows, concepts and not properties of the things that can be relational and non-relational. For instance the concept *being the same shape as David Lewis actually was at noon on January 1, 1970*, is a relational concept that presumably picks out an intrinsic property, namely a shape property. Whether they are valued magnitudes are intrinsic or extrinsic properties, is somewhat orthogonal question of whether it is best to pick them out by means of relational or non-relational concepts.

There is a further issue about the compatibility of Humean Supervenience with classical physics. This is a point that has been made well by Jeremy Butterfield (2006), and we can see the problem by looking at the different ways in which Lewis introduces Humean Supervenience.

> Humean Supervenience says that in a world like ours, the fundamental properties are local qualities: perfectly natural intrinsic properties of points, or of point-sized occupants of points. (Lewis, 1994a, 474)

Lewis goes back and forth between local properties and intrinsic properties of points here. These aren't the same thing. As Butterfield notes, 'local' is used in a few different ways throughout physics. One simple usage identifies local properties of a point with properties that supervene on intrinsic features of arbitrarily small regions around the point. To take an important example, the slope of a curve at a point may be a local property of the curve at that point without being intrinsic property of the point.

This raises a question: can we do classical physics with only intrinsic properties of points, and not even these further local properties? Butterfield argues, persuasively, that the answer is no. He notes, however, that there are some very mild weakenings of Humean Supervenience that avoid this difficulty. Here is a very simple one.

Call **Local Supervenience** the following thesis. For any length ε greater than 0, there is a length d less than ε with the following feature. All the facts about the world supervene on intrinsic features of objects and regions with diameter at most d, plus facts about the spatiotemporal arrangement of these objects and regions. This will mean that we can include all local qualities in the subvenient base, without assuming that these are intrinsic qualities of points. If the theory of intrinsicness in Weatherson (2006) is correct, we'll also be able to include vector-valued magnitudes in the subvenient base without assuming that these are intrinsic properties of points. (On my view, they will end up being intrinsic properties of asymmetrically shaped regions.) We still won't be able to accommodate entanglement relationships, but we will be able to capture classical physics. And, for the reasons discussed in the previous section, it would still be worthwhile to ask whether there are philosophical objections to Local Supervenience. A negative answer would greatly assist the arguments for compatibilism, and for nomological descriptivism.

Butterfield offers from theses like Local Supervenience to Lewis as friendly suggestions. But he thinks Lewis's focus on points and their properties would have led him to reject it. I don't want to get into the business of making counterfactual speculation about what Lewis would or would not have accepted. But I think he should have been happy to weaken Humean Supervenience to something like Local Supervenience. If the point of defending Humean Supervenience is not to defend its truth, but rather to assist in larger arguments for compatibilism, and for nomological descriptivism, then the big question to ask is whether a defence of Local Supervenience (against distinctively philosophical objections) would have served those causes just as well. And I think it's pretty clear that it would have. Showing that we have no philosophical reason to posit fundamental non-local features of reality would be enough to let us "resist philosophical arguments that there are more things in heaven and earth in physics has dreamt of" (Lewis, 1994a, 474). Lewis's work in defending Humean Supervenience has been invaluable to those of us who want to join this resistance. It wouldn't have been undermined if he'd allowed some local properties into the mix.

The Role of Naturalness in Lewis's Theory of Meaning

It is sometimes claimed (e.g., by Sider (2001a,b); Stalnaker (2004); Williams (2007); Weatherson (2003b)) that David Lewis's theory of predicate meaning assigns a central role to naturalness.[1] Some of the people who claim this also say that the theory they attribute to Lewis is true. The authors I have mentioned aren't as explicit as each other about exactly which theory they are attributing to Lewis, but the rough intuitive idea is that the meaning of a predicate is the most natural property that is more-or-less consistent with the usage of the predicate. Call this kind of interpretation the 'orthodox' interpretation of Lewis.[2] Recently Wolfgang Schwarz (2009, 209ff) has argued that the orthodox interpretation is a misinterpretation, and actually naturalness plays a much smaller role in Lewis's theory of meaning than is standardly assumed.[3] Simplifying a lot, one key strand in Schwarz's interpretation is that naturalness plays no role in the theory of meaning in Lewis (1969a, 1975b), since Lewis hadn't formulated the concept yet, and Lewis didn't abandon that theory of meaning, since he never announced he was abandoning it, so naturalness doesn't play anything like the role orthodoxy assigns to it.

In this article I attempt to steer a middle ground between these two positions. I'm going to defend the following parcel of theses. These are all exegetical claims, but I'm also interested in defending most of the theses that I ultimately attribute to Lewis, so getting clear on just what Lewis meant is of more than historical interest.

1. Naturalness matters to Lewis's (post-1983) theory of sentence meaning only insofar as it matters to his theory of rationality, and the theory of rationality matters to the (pre- and post-1983) theory of meaning.
2. Naturalness might play a slightly larger role in Lewis's theory of word meaning, but it isn't nearly as significant as the orthodox view suggests.
3. When we work through Lewis's theory of word and sentence meaning, we see that the orthodox interpretation assigns to Lewis a theory that isn't his theory of meaning, but is by his lights a useful heuristic.
4. An even better heuristic than 'meaning = use plus naturalness' would be 'meaning = predication plus naturalness', but even this would be a fallible heuristic, not a theory.
5. When correctly interpreted, Lewis's theory is invulnerable to the challenges put forward in Williams (2007).

I'm going to start by saying a little about the many roles naturalness plays in Lewis's philosophy, and about his big picture views on thought and meaning. Then I'll offer a number of arguments against the orthodox interpretation of Lewis's theory of sentence meaning. After

† Penultimate draft only. Please cite published version if possible. Final version published in *Journal for the History of Analytic Philosophy*, volume 1, number 10, .

[1] Holton (2003) is more nuanced, but does tell a similar story in the context of discussing Lewis's account of (potential) semantic indeterminacy. Weatherson (2010) follows Holton in this respect.

[2] As some further evidence for how orthodox the 'orthodox' interpretation is, note that Williams (2007) is a prize winning essay published with two commentaries in the *Philosophical Review*. That paper takes the orthodox interpretation as its starting point, and neither of the commentaries (Bays (2007) and Hawthorne (2007)) criticise this starting point.

[3] Schwarz (2006) develops his criticism of orthodoxy in more detail, and in English, but it is as yet unpublished.

that, I'll turn to Lewis's theory of word meaning, where it is harder to be quite clear about just what the theory is, and how much it might have changed once natural properties were added to the metaphysics. An appendix discusses some interpretative questions that arise if we are sceptical that any one division of properties can do all the work that Lewis has the natural/non-natural division do.

1 How Naturalness Enters The Theory of Meaning

Most of the core elements of David Lewis's philosophy were present, at least in outline, from his earliest work. The big exception is the theory of natural properties introduced in Lewis (1983b). As he says in that paper, he had previously believed that "set theory applied to possibilia is all the theory of properties that anyone could ever need" (Lewis, 1983b, 377n). Once he introduces this new concept of naturalness, Lewis puts it to all sorts of work throughout his philosophy. I'm rather sceptical that there is any one feature of properties that can do all the varied jobs Lewis wants naturalness to do, but the grounds for, and consequences of, this scepticism are a little orthogonal to the main theme of this paper, so I've set it aside.

As the orthodox interpretation stresses, Lewis has naturalness do some work in this theory of content. That he does think there's a connection between naturalness and content is undeniable from the most casual reading of his post-1983 work. But just how they are connected is less obvious. To spell out these connections, let's start with three Lewisian themes.

- Facts about linguistic meaning are to be explained in terms of facts about minds. In particular, to speak a language $\mathscr{L}$ is to have a convention of being truthful and trusting in $\mathscr{L}$ (Lewis, 1969a, 1975b). And to have such a convention is a matter of having certain beliefs and desires. So mental content is considerably prior to linguistic content in a Lewisian theory. Moreover, Lewis's theory of linguistic content is, in the first instance, a theory of *sentence* meaning, not a theory of *word* meaning.[4]
- The principle of charity plays a central role in Lewis's theory of mental content Lewis (1974a, 1994b). To a first approximation, a creature believes that p iff the best interpretation of the creature's behavioural dispositions includes the attribution of the belief that p to the creature. And, ceteris paribus, it is better to interpret a creature so that it is more rather than less rational. It will be pretty important for what follows that Lewis adopts a principle of charity that highlights *rationality*, not *truth*. It is also important to Lewis that we don't just interpret the individual creature, but creatures of a kind (Lewis, 1980a). I'm not going to focus on the social externalist features of Lewis's theory of mental states, but I think they assist the broader story I want to tell.
- Lewis's theory of mental content has it that mental contents are (what most of us would call) properties, not (what most of us would call) propositions (Lewis, 1979a). So a theory of natural properties can easily play a role in the theory of mental content, since

[4]These points are stressed by Wolfgang Schwarz (2006, 2009). He also notes that in "Putnam's Paradox" Lewis explicitly sets these parts of his theory aside so he can discuss Putnam's arguments on grounds most favourable to Putnam. As Schwarz says, this should make us suspicious of the central role "Putnam's Paradox" plays in defences of the orthodox interpretation. We will return to this point in the section on textual evidence for and against orthodoxy.

A referee notes, correctly, that the phrase 'in the first instance' is doing a lot of work here. That's right; we'll return in much more detail below to Lewisian theories of word meaning, and what role naturalness plays in them.

> mental contents are properties. If you think mental contents are propositions, the connection between naturalness and mental content will be more indirect. Just how indirect it is will depend on what your theory of propositions is. But if mental contents are Lewisian propositions, the connection may be very indirect indeed. After all, propositions that we might pick out with sentences containing words that denote very unnatural properties, such as *All emeroses are gred*, might be intuitively very natural.

Now let's see why we might end up with naturalness in the theory of meaning. An agent has certain dispositions. For instance, after seeing a bunch of green emeralds, and no non-green emeralds, in a large and diverse range of environments, she has a disposition to say "All emeralds are green". In virtue of what is she speaking a language in which "green" means green, and not grue? (Note that when *I* use "grue", I mean a property that only differs from greenness among objects which it is easy to tell that neither our agent, nor any of her interlocutors, could possibly be acquainted with at the time she makes the utterance in question.)

Let's say that $\mathscr{L}_1$ is English, i.e., a language in which "green" means green, and $\mathscr{L}_2$ a language which is similar to $\mathscr{L}_1$ except that "green" means grue. Our question is, what makes it the case that the agent is speaking $\mathscr{L}_1$ and not $\mathscr{L}_2$? That is, what makes it the case that the agent has adopted the convention of being truthful and trusting in $\mathscr{L}_1$, and not the convention or being truthful and trusting in $\mathscr{L}_2$?

We assumed that the agent has seen a lot of emeralds which are both green and grue. To a first approximation, it is more charitable to attribute to the agent the belief that all emeralds are green than the belief that all emeralds are grue because greenness is more natural than gruesomeness. As Lewis says, "The principles of charity will impute a bias towards believing things are green rather than grue" (1983b, 375). And for Lewis, charity requires imputing more reasonable interpretations. But why is it more charitable to attribute beliefs about greenness to beliefs about grueness? I think it is because we need more evidence to rationally form a belief that some class of things are all grue than we need to form a belief that everything in that class is green. And that's because, ceteris paribus, we need more evidence to rationally form a belief that all Fs are Gs than that all Fs are Hs when G is less natural than H. The agent has, we might assume, sufficient evidence to rationally believe that all emeralds are green, but not sufficient evidence to believe that all emeralds are grue.

So the first two Lewisian themes notes above, the reduction of linguistic meaning to mental content, and the centrality of a rationality-based principle of charity, push us towards thinking that naturalness is closely connected to mental content and hence to linguistic meaning. And it has pushed us towards thinking that if naturalness is connected to meaning, it is via this connection I've posited between naturalness and rational belief. Note that Lewis doesn't ever endorse anything like that general a connection, but I suspect he had something like this in mind when he wrote the sentence I quoted in the previous paragraph. We'll come back to this interpretative question at some length below.

But the argument I offered was a bit quick, because I ignored the third Lewisian theme: beliefs are relations to properties, not propositions. On Lewis's theory, to believe that all emeralds are green is to self-ascribe the property of being in a world where all emeralds are green. So if a certain body of evidence makes it possible for the agent to rationally believe that all emeralds are green, but not for her to believe that all emeralds are grue, and that's because rationality

is constitutively connected to naturalness, then that must be because the first of the following properties is more natural than the second:

- Being in a world where all emeralds are green
- Being in a world where all emeralds are grue

That could still be true, though it is notable how far removed we are from the intuitions that motivate the distinctions between more and less natural properties. It's not like there is some sense, intuitively, in which things that have the first property form a more unified class than things that have the second property.

So it's plausible that naturalness is connected to mental content, at least as long as naturalness is connected to rational belief. And since mental content is connected to linguistic content, we're now in the vicinity of the orthodox interpretation. But I don't think the orthodox interpretation can be right. I'll give four reasons for this, starting with the textual evidence for and against it.

2 Textual Evidence about Sentence Meaning

There is some *prima facie* textual evidence for the orthodox interpretation. But looking more careful at the context of these texts not just undermines the support the text gives to the orthodox interpretation, but actually tells against it. (This part of the paper is indebted even more than the rest to Wolfgang Schwarz's work, and could be easily skipped by those familiar with that work.)

I'll focus on the last seven pages of "New Work for a Theory of Universals". This is the part of "New Work" that uses the notion of naturalness, as introduced in the paper, to respond to Putnam's model-theoretic arguments for massive indeterminacy of meaning. Lewis actually responds to Putnam twice over. First, he responds to Putnam directly, by showing how adding naturalness to a use-based theory of sentence meaning avoids the 'just more theory' objection that's central to Putnam's argument. And when Lewis describes this direct response, he says things that sound a lot like the orthodox interpretation.

> I would instead propose that the saving constraint concerns the referent - not the referrer, and not the causal channels between the two. It takes two to make a reference, and we will not find the constraint if we look for it always on the wrong side of the relationship. Reference consists in part of what we do in language or thought when we refer, but in part it consists in eligibility of the referent. And this eligibility to be referred to is a matter of natural properties. (Lewis, 1983b, 371)

But after this direct response is finished, Lewis notes that he has conceded quite a lot to Putnam in making the response.

> You might well protest that Putnam's problem is misconceived, wherefore no need has been demonstrated for resources to solve it. … Where are the communicative intentions and the mutual expectations that seem to have so much to do with what we mean? In fact, where is thought? …I think the point is well taken, but

> I think it doesn't matter. If the problem of intentionality is rightly posed there will still be a threat of radical indeterminacy, there will still be a need for saving constraints, there will still be a remedy analogous to Merrill's suggested answer to Putnam, and there will still be a need for natural properties. (Lewis, 1983b, 373)

I noted earlier that Schwarz makes much of a similar passage in "Putnam's Paradox", and I think he is right to do so. Here's a crucial quote from that paper.

> I shall acquiesce in Putnam's linguistic turn: I shall discuss the semantic interpretation of language rather than the assignment of content to attitudes, thus ignoring the possibility that the latter settles the former. It would be better, I think, to start with the attitudes and go on to language. But I think that would relocate, rather than avoid, the problem; wherefore I may as well discuss it on Putnam's own terms. (Lewis, 1984b, 222)

That passage ends with a footnote where he says the final section of "New Work" contains a version of how the 'relocated' problem would be solved. So let's turn back to that. The following long portmanteau quote from pages 373 to 375 captures, I think, the heart of my interpretation.

> The problem of assigning content to functionally characterised states is to be solved by means of constraining principles. Foremost among these are principles of fit. …A state typically caused by round things before the eyes is a good candidate for interpretation as the visual experience of confronting something round; and its typical impact on the states interpreted as systems of belief ought to be interpreted as the exogenous addition of a belief that one is confronting something round, with whatever adjustment that addition calls for. …Call two worlds equivalent iff they are alike in respect of the subject's evidence and behaviour, and note that any decent world is equivalent inter alia to horrendously counterinductive worlds and to worlds where everything unobserved by the subject is horrendously nasty. …We can interchange equivalent worlds ad lib and preserve fit. So, given any fitting and reasonable interpretation, we can transform it into an equally fitting perverse interpretation by swapping equivalent worlds around …If we rely on principles of fit to do the whole job, we can expect radical indeterminacy of interpretation. We need further constraints, of the sort called principles of (sophisticated) charity, or of 'humanity'. [A footnote here refers to "Radical Interpretation".] Such principles call for interpretations according to which the subject has attitudes that we would deem reasonable for one who has lived the life that he has lived. (Unlike principles of crude charity, they call for imputations of error if he has lived under deceptive conditions.) These principles select among conflicting interpretations that equally well conform to the principles of fit. They impose *apriori* – albeit defeasible - presumptions about what sorts of things are apt to be believed and desired …**It is here that we need natural properties.** The principles of charity will impute a bias toward believing that things are green rather than grue …In short, they will impute eligible content …They will impute other things as well, but it is the imputed eligibility that matters to us at present. (Lewis, 1983b, 373-5, my emphasis)

I think that does a reasonably clear job of supporting the interpretation I set out in the introduction over the orthodox interpretation. Naturalness matters to linguistic meaning all right. But the chain of influence is very long and indirect. Naturalness constrains what is reasonable, reasonableness constrains charitable interpretations, charitable interpretations constrain mental content, and mental content constrains linguistic content. Without naturalness at the first step, we get excessive indeterminacy of content. With it, the Putnamian problems are solved. But there's no reason to think naturalness has any more direct role to play at any level in the theory of linguistic content.

In short, Lewis changed what he thought about rationality when he adopted the theory of natural properties. Since rationality was a part of his theory of mental content, and mental content determines linguistic content, this change had downstream consequences for what he said about linguistic content. But there wasn't any other way his theory of linguistic content changed, nor, contra orthodoxy, any direct link between naturalness and predicate meaning.

Moreover, when we look at the closest thing to a worked example in Lewis (1983b), we don't get any motivation for the orthodox interpretation. Here's the example he uses, which concerns mental content. Let f be any mapping from worlds to worlds such that the agent has the same evidence and behaviour in w and $f(w)$. Extend f to a mapping from sets of worlds to sets of worlds in the following (standard) way: $f(S) = \{f(w) : w \in S\}$. Then the agent's behaviour will be rationalised by her evidence just as much if she has credence function C and value function V, as if she has credence function C' and value function V', where $C'(f(S)) = C(S)$, and $V'(f(S)) = V(S)$. To relate this back to the familiar Goodmanian puzzle, let f map any world where all emeralds are green to nearest world where all emeralds are grue, and vice versa, and map any other world to itself. Then the above argument will say that the agent's behaviour is rationalised by her evidence just as much as if her credences are C as if they are C'. That is, her behaviour is rationalised by her evidence just as much if she gives very high credence to all emeralds being green as to all emeralds being grue. So understanding charity merely as rationalizing behaviour leaves us without a way to say that the agent believes unobserved emeralds are green and not grue.

Lewis's solution is to say that charity requires more than that. In particular, it requires that we assign natural rather than unnatural beliefs to agents where that is possible. I've argued above that this makes perfect sense if we connect naturalness with rationality. The crucial thing to note here is that this all happens a long time before we can set out the way that a sentence is used, since the way a sentence is used on Lewis's theory of linguistic content includes the beliefs that are formed on hearing it. So the discussion in "New Work" suggests that naturalness matters for content, but not in a way that can be easily factorised out. And that's exactly what I think is the best way to understand Lewis's theory.

3 Textual Evidence and Naturalness and Rationality

A major part of my argument above was that naturalness affected Lewis's theory of rationality. In particular, once he had naturalness to work with, he seemed to think that it was more rational to project natural rather than unnatural properties. The textual evidence for this is, I'll admit, fragmentary. But it is fairly widespread. Let's start with a quote we've already seen.

> The principles of charity will impute a bias toward believing that things are green rather than grue (Lewis, 1983b, 375)

As noted above, I assume this isn't a special feature of green and grue, but rather that there is a general principle in favour of projecting natural properties. But it would be good to have more evidence for that.

Lewis returns to the example of the believer in grue emeralds a few times. Here is one version of the story in *Plurality*.

> We think that some sorts of belief and desire ... would be unreasonable in a strong sense ... utterly unintelligible and nonsensical. Think of the man who, for no special reason, expects unexamined emeralds to be grue. ... What makes the perversely twisted assignment of content incorrect, however well it fits the subject's behaviour, is exactly that it assigns ineligible, unreasonable content when a more eligible assignment would have fit behaviour equally well. (Lewis, 1986b, 38-9)

And a little later, when replying to Kaplan's paradox, he says,

> Given a fitting assignment, we can scramble it into an equally fitting but perverse alternative assignment. Therefore a theory of content needs a second part: as well as principles of fit, we need 'principles of humanity', which create a presumption in favour of some sorts of content and against others. (Lewis, 1986b, 107)

He returns to this point again in "Reduction of Mind".

> [Folk psychology] sets presumptive limits on what our contents of belief and desire can be. Self-ascribed properties may be 'far from fundamental', I said – but not *too* far. Especially gruesome gerrymanders are *prima facie* ineligible to be contents of belief and desire. In short, folk psychology says that we make sense. It credits us with a modicum of rationality in our acting, believing and desiring. (Lewis, 1994b, 320 in reprint)

The running thread through these last three quotes is that our theory of mental content rules out gruesome assignments, and it does this because assigning rationality is constitutive of correctly interpreting. This can only work if naturalness is connected to rationality. I've attributed a stronger claim to Lewis, that not only is naturalness connected to rationality, but that the connection goes through projection.[5]

One piece of evidence for that is that Lewis says, in "Meaning Without Use" that Kripkenstein's challenge was "formerly Goodman's challenge" (Lewis, 1992, 109). He goes on to say that the solution to this challenge (or should that be 'these challenges') involves "carrying more baggage of primitive distinctions or ontological commitments than some of us might have hoped" (Lewis, 1992, 110). A footnote on that sentence cites "New Work", in case it isn't obvious that the baggage here is the distinction between natural and unnatural properties. So somehow, Lewis thinks that natural properties help solve Goodman's puzzle. I think that the simplest such solution is the right one to attribute to Lewis; natural properties are *prima facie* more eligible to be projected.

[5]The view I'm attributing to Lewis is endorsed by one prominent supporter of the orthodox interpretation, namely Ted Sider. See his (2012, 35ff).

A referee noted that this passage is a little odd; it appears to simply conflate a meta-semantical paradox with an epistemological paradox. But I think that just shows how much, for Lewis, meta-semantical questions are epistemological questions. Words get their meanings in virtue of our conventions. Our conventions consist of our beliefs and desires. And facts about rationality are, in part, constitutive of what we believe and desire.

Finally, consider the way in which the papers on natural properties are introduced in *Papers in Metaphysics and Epistemology*. Lewis says that "I had been persuaded by Goodman and others that all properties were equal: it was hopeless to try to distinguish 'natural' properties from gruesomely gerrymandered, disjunctive properties." (Lewis, 1999a, 1-2) A footnote refers to *Fact, Fiction and Forecast*. Of course, the point of "New Work" is that Lewis abandons this, explicitly Goodmanian, view. Now that he had learned property egalitarianism from Goodman of course doesn't show that once he became a property inegalitarian, he applied this to Goodman's own paradox. But it does seem striking that the only citation of an egalitarian view is of *Fact, Fiction and Forecast*. I take that to be some, inconclusive, evidence that Lewis did indeed think natural properties were related to Goodman's paradox.

Ultimately, it seems the textual evidence is this. There are many different occasions where Lewis makes clear there is a connection between naturalness and rationality, and in particular, between naturalness and the kind of rationality that is relevant to content assignment. There are hints that this connection goes via naturalness playing a role in solving Goodman's paradox. Notably, there is no other obvious way in which naturalness could connect to rationality. At least, I can neither think of another connection, nor see any evidence for another connection in the Lewis corpus. So I conclude, a little tentatively, that Lewis thought natural properties had a role to play in solving Goodman's paradox.

4 Word Meaning and Naturalness

In "Languages and Language", Lewis doesn't say that human linguistic practices merely determine truth conditions for the spoken sentences. That is, our linguistic practices don't merely determine which **language**, in Lewis's sense, we speak. They also determine, to some extent, a **grammar**, which specifies the truth conditional contribution of the various parts of the sentence. The grammar determines the "fine structure of meaning" (Lewis, 1975b, 177) of a sentence or phrase.

In comments on an earlier draft of this paper, an anonymous referee stressed that naturalness could enter directly into a theory of meaning once we stopped focussing on sentence meaning, and started looking on word meaning. I don't mean to say the referee was endorsing any particular role for naturalness in the theory of word meaning. But the point that we need to say more about the Lewisian approach to word meaning before we conclude that naturalness is only indirectly related to meaning is right. And I'm grateful for the encouragement to discuss it further.

Lewis has a short discussion of grammars in "Languages and Language", and another in "Radical Interpretation". It's worth looking at both of these in turn. I'll take "Languages and Language" first, since even though it has a slightly later publication date, in the respects we're discussing here it closely resembles the theory in *Convention*.

On pages 177-8 of that paper, Lewis notes three ways in which there may be indeterminacy in the grammar.

1. A subject's behavioural dispositions and anatomy might underdetermine their beliefs and desires.
2. The beliefs and desires might underdetermine the truth conditions of their language.
3. The truth conditions of the language might underdetermine the meanings of the individual words.

While Lewis does not think the second is actually a source of indeterminacy, he does think that the third is.

> My present discussion has been directed at the middle step … I have said … that the beliefs and desires of the subject and his fellows are such as to comprise a fully determinate convention of truthfulness and trust in some definite language. … I am inclined to share in Quine's doubts about the determinacy of the third step. (Lewis, 1975b, 178)

Lewis gives reasons for this inclination a few paragraphs earlier. He says that while we can say what it is for a community to speak one language rather than another, we can't say what it is for a community to speak one grammar rather than another. He says that we don't have any objective measures for evaluating grammars. And he says Quine's examples of indeterminacy of reference show that languages can have multiple good grammars, even if these disagree radically about the meaning of some constituents.

Notably, Lewis doesn't take to show that there is anything wrong with the notion of word meaning. He says it would be "absurd" (177) to conclude that. His conclusion here is more one of modesty rather than philosophical scepticism. We don't know how to extend the theory of sentence meaning he offers to a theory of word meaning, so we should do what we can without talking about word meaning.

The approach in "Radical Interpretation" has a bit more of a hint for how to restore semantic determinacy. The subject matter of that paper is how to solve for the mental and linguistic contents of a speaker, called Karl, given the physical facts about them. Lewis uses **M** for "a specification, in our language, of the meanings of expressions of Karl's language." (Lewis, 1974a, 333) He lists a number of constraints on a solution, including early versions of his principles of constitutive rationality. But the most notable constraint, from our perspective, is this:

> *The Principle of Generativity* constrains **M**: **M** should assign truth conditions to the sentences of Karl's language in a way that is at least finitely specifiable, and preferably also reasonably uniform and simple. (Lewis, 1974a, 339)

There's something very odd about this. Lewis, in 1974, didn't have a theory of what made an assignment simple. He needed his theory of natural properties to do that. Or, at least, once he had the theory of natural properties, it did all the work he ever wanted out of an account of simplicity.

Be that as it may, it does suggest that Lewis did think that simplicity of assignments could be used as a way of cutting down the third kind of semantic indeterminacy discussed in "Languages and Language". He doesn't think it would generate a fully determinate interpretation of Karl's language.

> It seems hopeless to deny, in the face of such examples as those in [Quine's "Ontological Relativity", pp. 30-39], that the truth conditions of full sentences in **M** do not sutfice to determine the rest of **M**: the parsings and the meanings of the constituents of sentences. At least, that is so unless there is something more than our Principle of Generativity to constrain this auxiliary syntactic and semantic apparatus. (Lewis, 1974a, 342-3)

It's notable that some of the examples Quine gives in "Ontological Relativity" are not cases where the alternative meanings are by any measure equally natural. This positive allusion to Quine's examples suggests a link to this comment in "Languages and Language"

> We should regard with suspicion any method that purports to settle objectively whether, in some tribe, "gavagai" is true of temporally continuant rabbits or time-slices thereof. You can give their language a good grammar of either kind—and that's that. (Lewis, 1975b, 177)

Note that he doesn't say 'equally' good. And note also how this contrasts with the attitude he takes towards the prospects of indeterminacy in sentence meaning. I earlier quoted him saying that part of the point of "Languages and Language" was to show how the second type of indeterminacy didn't arise. He ends "Radical Interpretation" with this 'credo'.

> Could indeterminacy of beliefs, desires, and truth conditions also arise because two different solutions both fit all the constraints perfectly? Here is the place to hold the line. This sort of indeterminacy has not been shown by convincing examples, and neither could it be shown–to me–by proof. *Credo*: if ever you prove to me that all the constraints we have yet found could permit two perfect solutions, differing otherwise than in the auxiliary apparatus of **M**, then you will have proved that we have not yet found all the constraints. (Lewis, 1974a, 343)

So that's where things stood before 1983. Lewis thought he had a theory that eliminated, or at least minimised, indeterminacy at the level of truth conditions. But he didn't think his theory eliminated indeterminacy, even quite radical indeterminacy, in word meanings. And he didn't seem bothered by this aspect of the theory; indeed, he thought Quine's arguments showed that we shouldn't eliminate this kind of indeterminacy.

This attitude towards Quinean arguments for indeterminacy is obviously a striking contrast to the forcefulness, and rapidity, with which he responded to Putnam's arguments for indeterminacy. That shouldn't be too surprising once we attend to Lewis's threefold distinction between kinds of indeterminacy. Quine was arguing that indeterminacy of the third kind was rampant. Putnam was arguing that indeterminacy of the second kind was rampant. And, as Lewis announced in "Radical Interpretation", he wasn't going to believe any such argument.

Still, we might wonder whether the resources he brought to bear in responding to Putnam also help respond to Quine. Or, perhaps more importantly for exegetical reasons, we might wonder whether Lewis thought they were useful in responding to Quine. The evidence from "New Work" seems to suggest a negative answer to the latter question. Lewis never says that one of the things you can do with the distinction between natural and unnatural properties is respond to arguments for Quinean indeterminacy. And that's despite the fact that "New

Work" has a very survey-like feel; the bulk of the paper is a long list of philosophical work that a theory of universals can do.

In "Putnam's Paradox" there is a brief footnote on Quine's arguments for indeterminacy. It reads

> It is not clear how much indeterminacy might be expected to remain. For instance, what of Quine's famous example? His rabbit-stages, undetached rabbit parts, and rabbit-fusion seem only a little, if any, less eligible than rabbits themselves. (Lewis, 1984b, 228n)

As I've stressed repeatedly, following Schwarz, taking the disclaimers at the start of "Putnam's Paradox" seriously means that we have to be careful in interpreting what Lewis says about how words acquire determinate meaning in that paper. But even before we adjust for the disclaimers, this is hardly a ringing rejection of Quine's indeterminacy arguments. The contrast to Lewis's attitude towards Putnam's arguments is striking. Since it is the very same contrast that we saw in both "Languages and Language" and "Radical Interpretation", I think it is fair to assume that he continued to think Quine's arguments were considerably stronger than Putnam's.

But there is, perhaps, a change of view in "Meaning Without Use". Here's the problem Lewis addresses at the end of that paper. Let $\mathscr{L}_1$ once again be English as we currently understand it, and let $\mathscr{L}_3$ be just like English, except that it doesn't assign any truth conditions to sentences over a thousand words long.[6] Do our actual linguistic practices manifest a convention of trust in $\mathscr{L}_1$, or trust in $\mathscr{L}_3$? Lewis argues that it is more like a convention of trust in $\mathscr{L}_3$. If someone utters a very long sentence, we expect some kind of performance error, at best. We don't, in general, believe what they say. So the theory of "Languages and Language" seems to predict that these long sentences have no truth conditions. But that's wrong, so the theory must be corrected.

Lewis's correction appeals, it seems, to natural properties in fixing a grammar. He says that linguistic practice determines truth conditions for a fragment of the language that is widely used. Those truth conditions determine meanings of words. This determination requires natural properties; without them the Quinean problems multiply indefinitely. We then use those word meanings to determine the meaning of unused sentences. A long footnote suggests that the procedure might not be restricted to unused sentences. As long as there is a large enough fragment in which there are conventions of truthfulness and trust, we can extrapolate from that to other parts of the language that are used.

This is a marked deviation from anything Lewis had said until then. From the earliest writings, he had stressed a step-by-step approach to content determination. Behavioural dispositions plus physical and biological constraints determine mental content; mental content determines sentence meaning; and sentence meaning determines word meaning. In "Meaning Without Use", it seemed the last two steps were being somewhat merged.

But we shouldn't overstate how much the third step was allowed to encroach on the second. Lewis does think we need to rule out 'bent' grammars, which don't assign any truth conditions to sentences over a thousand words long, or which give sentences different meanings to what we'd expect if the word 'cabbage' appears forty times. But he doesn't think we need to rule out

[6] If you think sentences with a thousand words are too easy to understand for the argument of this paragraph, make the threshold higher; as long as the threshold is finite, it won't affect the argument.

any 'straight' grammar, which includes "any grammar that any linguist would actually propose." (Lewis, 1992, 109)

So Lewis's focus here is to rule out unnatural *compositional rules*, not unnatural assignments of content to individual words. The reference to linguists here might be useful. Linguists tend to spend much more time on compositional rules than they do on the contents on individual predicates. Notably, Quine didn't argue for indeterminacy by positing indeterminacy in the compositional rules of the language; his non-standard interpretations all share a standard syntax. If we posit that Lewis thought that there was little syntactic indeterminacy in the language, like there is little indeterminacy at the level of truth conditions of sentences, we can tell a story that doesn't involve too many unsignalled changes of view. Here's how I would tell that story in some more detail.

Lewis's early view, expressed clearly in "Radical Interpretation" and "Languages and Language", and not retracted before, I think, 1992, has the following parts:

1. Conventions of truthfulness and trust determine (very sharply) truth conditions for sentences in a speaker's language.
2. Any reasonably good grammar, i.e., assignment of word meanings and compositional rules, that is consistent with the truth conditions is not determinately wrong. There is potentially substantial indeterminacy in the meaning of any given word, because there are many reasonably good grammars consistent with the truth conditions.

After 1983, 'simplicity' was understood in terms of naturalness, but otherwise the story doesn't change a lot.

The later view, which goes by somewhat more quickly in "Meaning Without Use", has the following parts:

1. Conventions of truthfulness and trust in (the bulk of) the used fragment of the language determine truth conditions for that fragment.
2. Naturalness considerations determine the compositional rules for the language by extrapolation from that grammar.
3. Word meanings are determined, so far as they are determinate, by the truth conditions for sentences, plus the compositional rules.
4. Truth conditions for sentences outside the used fragment are determined by the word meanings and the compositional rules.

Neither of these views look much like the orthodox view. Remember that the orthodox view has it that considerations of naturalness can be used to resolve debates in metaphysics. That's certainly the use that Sider (2001a) makes of the orthodox view. But on the early view, simplicity considerations only come in after the truth conditions for every sentence have been determined, and hence so that all debates are settled. And on the later view, simplicity considerations primarily are used to settle truth conditions for unused, or at least unusual, sentences.

Now if you thought the salient fragment in point 1 of the later view was small, and if you thought naturalness had a major role to play in step 3 of the later view, you would get back to something like the orthodox view. But I don't see the textual evidence for either of those positions. Lewis says that "the used fragment is large and varied." (Lewis, 1992, 110) It doesn't look like he is positing wholesale changes to his view on the determination of truth conditions.

He is positing some changes; the last two pages of the paper are clearly marked as deviations from his earlier position. But both the examples he uses and the rhetoric around them suggests that the bulk of the changes happen at point 2. Naturalness considerations constraint the syntax of a language much more tightly than they constrain the assignment of meaning to a given word. In sum, at no point in the evolution of his views did Lewis seem to endorse the orthodox interpretation, even as a theory of word meaning.

5 An Argument for the Orthodox Interpretation

So far I've argued that there is no solid textual support for the orthodox interpretation. My rival interpretation relied on there being a connection between naturalness and induction, and as we've just seen, there is some textual evidence for this. But perhaps there is a more indirect way to motivate the orthodox interpretation of Lewis. The orthodox interpretation attributes to Lewis a theory that is quite attractive as a theory of semantic determinacy and indeterminacy. Call that theory the **U&N Theory**, short for the **U**se plus **N**aturalness theory of meaning. Since Lewis was clearly looking for such a theory when he discussed naturalness in the context of his theory of content, it is reasonably charitable to attribute the **U&N Theory** to him, as the orthodox interpretation does.

My response to this will be in three parts. First, I'll argue in this section that my rival interpretation attributes to Lewis a theory of semantic determinacy and indeterminacy that does just as well at capturing the facts Lewis wanted a theory to capture, so there's no charity based reason to attribute the **U&N Theory** to him (And, as we saw in the previous section, there's no direct textual reason to attribute it to him either.) Second, the **U&N Theory** is subject to the criticisms in Williams (2007), while the theory I attribute to Lewis is not. Third, the **U** part of the **U&N Theory** is hopelessly vague; it isn't clear how to say what 'use' is on a Lewisian theory that makes it suitable to add to naturalness to deliver meanings. Either use is so thick that naturalness is unneeded, or it is so thin that naturalness won't be sufficient to set meaning. So actually it isn't particularly charitable to attribute this theory to him.

Still, let's start with the attractions of the **U&N Theory**. On the one hand, agents are inclined to say "All emeralds are green" both in situations where they've seen a lot of green emeralds (and no non-green ones) and in situations where they've seen a lot of grue emeralds (and no non-grue ones). That's because, of course, those are exactly the same situations. So at first glance, it doesn't look like the way in which "green" is used will determine whether it means green or grue. On the other hand, once we add a requirement that terms have a relatively natural meaning, we do get this to fall out as a result. Moreover we can even see how this falls out of a recognisably Lewisian approach to meaning.

Consider again our agent who says "All emeralds are green" after seeing a lot of emeralds that are both green and grue. And remember that for her to speak a language, she must typically conform to conventions of truthfulness and trust in that language. Now if the agent was speaking $\mathscr{L}_2$, she would have to think that she's doing an OK job of being truthful in $\mathscr{L}_2$ by saying "All emeralds are green". But that would be crazy. Why should she think that all emeralds are grue given her evidence base? To attribute to her that belief would be to gratuitously attribute irrational beliefs to her. And on Lewis's picture, gratuitous attributions of irrationality are false. So the agent doesn't have that belief. So she's not speaking $\mathscr{L}_2$.

Things are even clearer from the perspective of hearers. A hearer of "All emeralds are green" would be completely crazy to come to believe that all emeralds are grue. The hearer knows, after all, that the speaker has no acquaintance with the emeralds that would have to be blue for all emeralds to be grue. So the hearer knows that this utterance could not be sufficient evidence to believe that all emeralds are grue. Yet if she speaks $\mathscr{L}_2$, she is disposed to believe that all emeralds are grue on hearing "All emeralds are green". She isn't irrational, or at least we shouldn't assign irrationality to her so quickly, so she doesn't speak $\mathscr{L}_2$.

So it looks like in this one case at least, we have a case where use plus naturalness gives us the right theory. Agents are disposed to use "green" to describe emeralds that are green/grue. But the fact that greenness is more natural than gruesomeness makes it more appropriate to attribute to them a convention according to which "All emeralds are green" means that all emeralds are green and not that all emeralds are grue.

But more carefully, what we should say is that the **U&N Theory** gives us the right result in this case. It doesn't follow that it will work in all cases, or anything like it. And it doesn't follow that it works for the right reasons. As we'll see, neither of those claims are true. In fact, just re-reading the last three paragraphs should undermine the second claim. Because we just saw a derivation that the agents are not speaking $\mathscr{L}_2$, that didn't even appeal to the **U&N Theory**. Rather, that derivation simply used the theory of meaning in *Convention* and the theory of mental content in "Radical Interpretation". It's true that the latter theory assigns a special role to rationality, and the theory of rationality we used has, among other things, a role for natural properties, but that is very different to the idea that naturalness feeds directly into the theory of meaning in the way the orthodox interpretation says. As I said at the start, I think the best interpretation of Lewis is that he changed his theory of *rationality* in 1983, but that's the only change to his theory of *meaning*.

Put another way, these reflections on "green" and "grue" are consistent with the view that the **U&N Theory** is a false *theory*, but a useful *heuristic*. It's a useful heuristic because it agrees with the true Lewisian theory in core cases, and is much easier to apply. That's exactly what I think the **U&N Theory** is, both as a matter of fact, and as a matter of Lewis interpretation.

6 Indeterminacy and Radically Deviant Interpretations

If the **U&N Theory** is a heuristic not a theory, we should expect that it will break down in extreme cases. That's exactly what we see in the cases discussed in Williams (2007). Those cases highlight the fact that a Lewisian theorist needs to be careful that we don't end up concluding that normal people, such as the agent in our example who says "All emeralds are green", speak $\mathscr{L}_4$. $\mathscr{L}_4$ is a language in which all sentences express claims about a particular mathematical model (essentially a Henkin model of the sentence the agent accepts), and it is set up in such a way that ordinary English sentences come out true, and about very natural parts of the model. On the **U&N Theory**, it could easily turn out that ordinary speakers are speaking $\mathscr{L}_4$, since the assigned meanings are so natural. We can see this isn't a consequence of *Lewis's* theory by working through the case from first principles. I have two arguments here, the first of them relying on some slightly contentious claims about the epistemology of mathematics, the second less contentious.

Assume, for reductio, that ordinary speakers are speaking $\mathscr{L}_4$. So, for instance, when O'Leary says "The beer is in the fridge", what he says is that a certain complicated mathematical model has a certain property. (And indeed it has that property.) Now this won't be a particularly rational thing for O'Leary to say unless he knows more mathematics than ordinary folks like him ordinarily do. So if O'Leary has adopted a convention of truthfulness and trust in $\mathscr{L}_4$, then uttering "The beer is in the fridge" would be irrational, even if he is standing in front of the open fridge, looking at the beer. That's a gratuitous assignment of irrationality, and gratuitous assignments of irrationality are false, so O'Leary doesn't speak $\mathscr{L}_4$.

Perhaps that is too quick. After all, the mathematical claim that $\mathscr{L}_4$ associates with "The beer is in the fridge" is a necessary truth. And Lewis's theory of content is intentional, not hyper-intentional. So O'Leary does know it is true. (And when he is standing in front of the fridge, there's even a sense that he knows that "The beer is in the fridge" expresses a truth, if $\mathscr{L}_4$ is really his language.) I think that's probably not the right sense of "rational", and I'm not altogether sure how much hostility to hyper-intensionalism we should attribute to Lewis. But so as to avoid these questions, it's easier to consider a different argument that focusses attention on O'Leary's audience.

When O'Leary says "The beer is in the fridge", Daniels hears him, and then walks to the fridge. Why does Daniels make such a walk? Well, he wants beer, and believes it is in the fridge. That looks like a nice rational explanation. But why does he believe the beer is in the fridge? I say it's because he's (rationally) adopted a convention of truthfulness and trust in $\mathscr{L}_1$, and so he rationally comes to believe the beer is in the fridge when O'Leary says "The beer is in the fridge". On the assumption that O'Leary and Daniels speak $\mathscr{L}_4$, none of this story goes through. But we must have some rational explanation of why O'Leary's statement makes Daniels walk to the fridge. So O'Leary and Daniels must not be speaking $\mathscr{L}_4$.

Michael Morreau pointed out (when I presented this talk at CSMN) that the preceding argument may be too quick. Perhaps there is a way of rationalising Daniels's actions upon hearing O'Leary's words consistent with the idea that they both speak $\mathscr{L}_4$. Perhaps, for instance, Daniels's walking to the fridge constitutes saying something in a complicated sign language, and that thing is the rational reply to what O'Leary said. If this kind of response works, and I have no reason to think it won't, the solution is to increase the costs to Daniels of performing such a reply. For instance, not too long ago I heard Mayor Bloomberg say "Lower Manhattan is being evacuated because of the impending hurricane", and I (and my family) packed up and evacuated from Lower Manhattan. Even if one could find an interpretation of our actions in evacuating that made them constitute the assertion of a sensible reply to Bloomberg's mathematical assertion in $\mathscr{L}_4$, it would be irrational to think I made such an assertion. Evacuating ahead of a storm with an infant is not fun - if it was that hard to make mathematical assertions, I wouldn't make them! And I certainly wouldn't make them in reply to someone who wouldn't even see my gestures. So I think at least some of the actions that are rationalised by testimony, interpreted as sentences of $\mathscr{L}_1$, are not rationalised by testimony, interpreted as $\mathscr{L}_4$. By the kind of appeal to the principle of charity we have used a lot already, that means that $\mathscr{L}_4$ is not the language most people speak.

The central point here is that when we are ruling out particularly deviant interpretations of some speakers, we have to make heavy use of the requirement that the interpretation of their shared language rationalises what they do. In part that means it must rationalise why they utter the strings that they do in fact utter. And when we're considering this, we should remember

the role of naturalness in a theory of rationality. But it also means that it must rationalise why people respond to various strings with non-linguistic actions, such as walking to the fridge, or evacuating Lower Manhattan. Naturalness has less of a role to play here, but the Lewisian theory still gets the right answers provided we apply it carefully. Since the Lewisian theory gets the right answers, and the **U&N Theory** gets the wrong answers, it follows that the **U&N Theory** isn't Lewis's theory, and so orthodoxy is wrong.

7 What is the Use of a Predicate?

We concluded the last section with an argument that Lewis isn't vulnerable to the claim that his theory assigns complicated mathematical claims as the meanings of ordinary English sentences. That interpretation, we argued, is inconsistent with the way those sentences are used. In particular, it is inconsistent with the way that *hearers* use sentences to guide their actions.

So far so good, we might think. But notice how much has been packed into the notion of use to get us this far. In identifying the use O'Leary makes of "The beer is in the fridge", we have to say a lot about O'Leary's beliefs and desires. And in identifying the use Daniels makes of it, we *primarily* talk about the sentence's effects on Daniels's beliefs and desires. That is, just saying how the sentence is used requires saying a lot about mental states of speakers. And that will often require appealing to constitutive rationality; we say that Daniels's beliefs about the fridge changed because we need to rationalise his fridge-directed behaviour.

And this should all make us suspicious about the prospects for identifying meaning (in a Lewisian theory) with use plus naturalness. The argument above that naturalness mattered to meaning relied on the idea that naturalness matters because it affects which states are rational, and hence which states are actualised. A belief that all emeralds are grue is unnatural, so it is hard to hold. And since it is hard to hold, it is hard to think one is conforming to a convention of truthfulness in a language if one utters sentences that mean, in that language, that all emeralds are grue. That's why it is wrong, *ceteris paribus,* to interpret people as speaking about grueness.

But now consider what happened when we were talking about Daniels and O'Leary. Even to say how they were using the sentence "The beer is in the fridge", we had to say what they believed before and after the sentence was uttered. In other words, their mental states were constitutive of the way the sentence was used. Now add in the extra premise, argued for above, that naturalness matters to Lewis's theory of linguistic content because, and only because, it matters to his theory of mental content. (And it only matters to mental content because it matters to the principle of charity that Lewis uses.) If mental states, and their changes, are part of how the sentences are used, it will be rather misleading to say that meaning is determined by use plus naturalness. A better thing to say is that meaning is determined by use, and that some key parts of use, i.e., mental states of speakers and hearers, are determined in part by naturalness.

So I'm sceptical of the **U&N Theory**. We can put the argument of the last few paragraphs as a dilemma. There are richer and thinner ways of identifying the use to which a sentence is put. A thin way might, for instance, just focus on the observable state of the part of the physical world in which the sentence is uttered. A rich way might include include, inter alia, the use that is made of the sentence in the management of belief and the generation of rational action. If we adopt the thin way of thinking about use, then adding naturalness won't be enough to say

what makes it the case that O'Leary and Daniels are speaking $\mathscr{L}_1$ rather than $\mathscr{L}_4$. If we adopt the rich way of thinking about use, then the role that naturalness plays in the theory of meaning has been incorporated into the metaphysics of use. Neither way makes the **U&N Theory** true while assigning naturalness an independent role. This dilemma isn't just an argument that we shouldn't attribute the **U&N Theory** to Lewis; it is an argument against anyone adopting that theory.

8 From Theory to Applied Semantics

So far we've argued that Lewis's semantic theory did not look a lot like the orthodox interpretation. It's true that he thought the way a sentence was used was of primary importance in determining its meaning. And it's true that he thought naturalness mattered to meaning. But that wasn't because naturalness came in to resolve the indeterminacy left in a use-based theory of meaning. Rather, it was because naturalness was in a part of the theory of mental content, and specifying the mental states of speakers and hearers is part of specifying how the sentence is used.

But note that these considerations apply primarily to investigations at a very high level of generality, such as when we're trying to solve the problems described in "Radical Interpretation". They don't apply to investigations into applied semantics. Let's say we are trying to figure out what O'Leary and Daniels mean by "green". And assume that we are taking for granted that they are speaking a language which is, in most respects, like English. This is hardly unusual in ordinary work in applied semantics. If we are writing a paper on the semantics of colour terms, a paper like, say, "Naming the Colours", we don't concern ourselves with the possibility that every sentence in the language refers to some complicated mathematical claim or other.

Now given those assumptions, we can identify a moderately thin notion of use. We know that O'Leary uses "green" to describe things that are, by appearance, both green and grue. We also know that when O'Leary makes such a description, Daniels expects the object will be both green and grue. So focus on a notion of use such that the *use* of a predicate just is a function of which objects speakers will typically apply the predicate to, and which properties hearers take those objects to have once they hear the predication. If we wanted to be more precise, we could call this notion of 'use' simply *predication*. When we are doing applied semantics, especially when we are trying to figure out the meaning of predicates, we typically know which objects a speaker is disposed to predicate a predicate of, and that's the salient feature of use. (This is why I said the most accurate heuristic would be meaning is predication plus naturalness; predication is the bit of use we care about in this context.)

This identification of use wouldn't make any sense if we were engaged in theorising at a much more abstract level. If we are doing radical interpretation, then we have to take non-semantic inputs, and solve simultaneously for the values of the subject term and the predicate term in a (simple) sentence. But when we are just doing applied semantics, and working just on the meaning of a term like "green" in a well-functioning language, we can presuppose facts about the denotation of the subject term in sentences like *S is green*, and presuppose facts about what is the subject and what is the predicate in that sentence, and then we can look at which properties hearers come to associate with that very object on hearing that sentence.

Now that we have a notion of use that's distinct from naturalness, we can ask whether it is plausible that predicate meaning is use (in that sense) plus naturalness. And, quite plausibly, the answer is yes. The arguments in Sider (2001a) and Weatherson (2003b) in favour of this theory look like, at the very least, good arguments that the theory does the right job in resolving Kripkensteinian problems. The theory is immune to objections based on radical reinterpretations of the language, as in Williams (2007), because those will be inconsistent with the use so defined. And the theory fits nicely into Lewis's broader theory of meaning, i.e., his metasemantics, which is in turn well motivated. So I think there are good reasons to hold that when we're doing applied semantics, the **U&N Theory** delivers the right verdicts, and delivers them for Lewisian reasons. That's the heart of what's true about the **U&N Theory**, even if it isn't a fully general theory of meaning.

Centrality and Marginalisation

Brian Weatherson

1 Welcome to the History of Late Analytic Philosophy

It's a good time to be doing history of late analytic philosophy. There is flurry of new and exciting work on how philosophy got from the death pangs of positivism and ordinary language philosophy to where it is today. Some may see this as a much needed gap in the literature. Indeed, there are a couple of reasons for scepticism about there being such a field as history of late analytic philosophy, both of which are plausible but wrong.

One reason is that it is too recent. But it can't be too recent for general historical study; there are courses in history departments on September 11, so it's not like looking at philosophy from thirty to forty years ago is rushing in where historians fear to tread. And indeed, if logical positivism could be treated historically in the 1960s, and ordinary language philosophy could be treated historically at the turn of the century, it seems a reasonable time to look back at the important works of the 1970s that established the contemporary era in philosophy.

Another reason is that we all know it so well. We are still so engaged with the key works by Kripke, Lewis, Burge, Perry, Thomson and so on that we don't need to also look at them the way we look at Descartes, Locke and Hume. But this, it turns out, is not true. Books by Daniel Nolan (2005) and Wolfgang Schwarz (2009) changed the way that some philosophers, even those who knew the Lewisian corpus fairly well, changed the way they read Lewis. There has also been a minor flurry of work on how important the Gödel/Schmidt case is to the argument of *Naming and Necessity* (Devitt, 2010; Ichikawa et al., 2012; Machery et al., 2012).

But that's nothing compared to the bombshell that is *Philosophy Without Intuitions*. (Cappelen 2012; all page citations, unless otherwise noted, to this book.) Herman Cappelen shows, extremely convincingly to my eyes at least, that intuitions play a much smaller role in late analytic philosophy than many philosophers thought. Indeed, there is a lot of textual evidence both for the claim that intuitions don't do much philosophical work, and for the claim that many people have said that they do. The first of these claims is all to the good, says Cappelen, since there isn't a particularly good epistemological defence of the use of intuitions.

The evidence for Cappelen's claims comes in two parts. The first part, which I won't discuss much here, is an extended argument that words like 'intuitively', or 'counterintuitive', as they appear in philosophical discourse, don't in general function to pick out, or even draw attention to, any distinctive kind of mental state we could call an 'intuition'. The second part argues that when we look at the actual introduction of thought experiments into late analytic philosophy, we don't see the appeal to intuitions that many philosophers seem to think go along with thought experiments. Rather, we see a whole host of interesting philosophical moves. Sometimes a thought experiment functions to highlight an explanandum. Sometimes it gives us a *prima facie* plausible thesis that we then argue for (or against) at great length. Sometimes it just raises a puzzle.

One upshot of this historical work, one that Cappelen I think does a good job highlighting, is that contemporary philosophy is much more *interesting* than its practitioners sometimes take

† Unpublished. Draft of paper commissioned for symposium on Herman Cappelen's *Philosophy Without Intuitions*. Thanks to Herman Cappelen and Ishani Maitra for many discussions about the material in this paper.

it to be. Philosophy is a way of investigating hard questions about the world, often at great expense in terms of human capital, but with thankfully little in the way of other expenses. It isn't a matter of tidying up conceptual space. Thinking of philosophy this way should, I think, help us see why so many different kinds of projects are philosophically important.

2 Centrality and Its Discontents

The big goal of Cappelen's book is to refute the view, which he dubs Centrality, that intuitions (of a certain kind) are central to analytic philosophy, and in particular that they are a primary source of evidence for analytic philosophers. The intuitions that he has in mind have these three characteristics. (The quotes are from pages 112-3, where these features are articulated.)

F1: Phenomenology "An Intuitive Judgment has a distinctive phenomenology" .

F2: Rock "An intuitive judgment has a special epistemic status …Intuitive judgments justify, but they need no justification".

F3: Conceptual A judgment is an intuition "only if it is justified solely by the subjects' conceptual competence".

There's some more detail on F2, but we'll get to that in section 6. And there's a fourth characteristic of intuitions that I want to add.

F4: Speed Intuitions are rapid reactions..[1]

I'm going to spend much of this paper defending a view that intuitions characterised by F2 and F4 do play a role, though perhaps not a *central* role, in philosophy. But I do think that intuitions characterised by F1 and F3 are just not important to philosophy. Indeed, I think it's a very important fact that they are not that important.

The claim that intuitions have a distinctive phenomenology is mostly harmless but, it seems to me, false. I certainly don't find anything in common when I introspect my judgments that, say, no set is a member of itself, or that losing a limb would seriously reduce my happiness, or that the only language I think in is English. It will fall out of the view I'm defending that the best intuitions have no phenomenology, but I don't think that's a particularly important fact about them.

But the claim that intuitions derive solely from conceptual competencies, plus the claim that these are the central source of evidence in philosophy, is both wrong and dangerous. If that conjunction were true, we'd expect most philosophical conclusions to be conceptual truths (whatever those are). I'm not going to take a stand on whether there are conceptual truths, but I think it is pretty obvious that conceptual truths won't help much resolve the following debates. (Compare the list E1-E6 on pages 200-201, which I'm basically just extending.)

- Do bans on pornography involve trading off speech rights versus welfare considerations, or do they just involve evaluating the free speech interests of different groups?
- Is it permissible to eat whales?

[1] My own views about the importance of this, as well as much else in this paper, owe a lot to Jennifer Nagel (2007, 2013).

- Under what circumstances is it permissible to end a terminally ill patient's life, or to withhold life-saving treatment?
- Is all context dependency in language traceable to the presence of bindable variables?
- Does belief have a phenomenology?
- Which animals (and which non-animals) have beliefs?

If philosophy uses largely conceptual evidence, these aren't philosophical questions. More generally, if Centrality (in Cappelen's sense) is true of philosophy, then feminist philosophy, legal philosophy, political philosophy, bioethics, philosophy of language and (most of) philosophy of mind are not part of philosophy. (This list is far from exhaustive; making philosophy Centrality-friendly would involve writing out huge swathes of the discipline.)

Modus tollens obviously beckons. But as Cappelen notes (213), one occasional reaction to this is to identify certain parts of philosophy as the 'Core' of the discipline, and say Centrality is true of those. If Centrality is true of the core of philosophy, then feminist philosophy *etc*., are not part of the core of the field. Maybe now some people would be disposed to use modus ponens not modus tollens.

That would be a large mistake. It would have shocked Plato, and Locke, and Hume, and practically every other major figure in the history of philosophy to learn that political philosophy wasn't central to the field. I do think (contra some of what Cappelen says) that some philosophy involves a priori and conceptual investigation. Indeed, I even do some of it. But it's not true that when I'm doing that I'm doing work that's deeper, or more philosophical, or more central to philosophy than the work that, for example, Rae Langton or Susan Moller Okin or Tamar Szabó Gendler or Sarah-Jane Leslie do.

This reason alone suffices for me to hope that Cappelen's book has a very wide readership. Centrality isn't true, but it is I think widely believed to true of at least some parts of the field. (Cappelen quotes many people endorsing this view.) I suspect that on the basis of this mistake, the parts of philosophy about which Centrality is not obviously false (especially metaphysics and epistemology) have been seen as more central to the discipline than they really ought to be. That's not a bad state of affairs for metaphysicians and epistemologists, but it's not good for philosophy, and I hope that Cappelen's book helps put a stop to it.

3 Intuitions in Detective Work

Despite my very broad sympathy with Cappelen's project, I do think there's a role for intuitions of some kind in philosophy. Just what this kind is, and what this role is, will take some spelling out to avoid Cappelen's arguments. So that's what I'll do for the next few pages.

The intuitions I have in mind are characterised by F2 and F4; they are default justified, and they are fast. Here's how I think these kinds of intuitions could matter philosophically.

When humans are growing up, they develop a lot of cognitive skills. Some of these skills are grounded in specific bits of propositional knowledge. We learn to count in part by learning that 2 comes after 1, and 3 comes after 2, and so on. But not all of them are. We learn how to tell causation from correlation, at least in simple cases, by developing various heuristics, none of which come close to a full theory of causation. Indeed, none of these heuristics would even be true, if stated as universal generalisations. But this ability to pick out which of the many

predecessors of an event is its cause is one we develop very early (Gopnik, 2009, 33-44), and it is vital to navigating the world.

I think we develop a lot of skills like that; skills which either go beyond our propositional knowledge, or at the very least are hard to articulate in terms of propositions. That we have these kinds of skills should hardly be news to philosophers; under the label 'heuristics' they have become quite familiar thanks to the work of, among others, Daniel Kahneman. They occasionally get a bad press, because one central way in which psychologists detect them is by seeing where they lead to errors that careful thought would correct. (For instance, our heuristics sometimes say that a conjunction is more probable than one of the conjuncts, and careful thinking would correct this.) But this should not blind us to the fact that these incredibly fast heuristics are often very reliable; reliable enough to be an independent check on our theorising.

The use of the term 'intuition' to pick out these heuristics isn't particularly idiosyncratic; Kahneman (2011) himself moves back and forth freely between the two terms. He approvingly cites Herbert Simon's remark that "intuition is nothing more and nothing less than recognition", which I think is basically right. We intuit that a is F by recognising that it has the tell-tale signs of Fhood. Of course we're a million miles from conceptual or *a priori* reasoning here; as I said, I agree entirely with Cappelen that F3 is not a feature of any philosophically significant source of evidence. Here are a couple of cases, one real life and one fictional, that draw out far removed intuitive thinking can be from a priori or conceptual thinking. The first is from Kahneman's description of a case reported by Gary Klein (1999); the second is from (Norwegian) crime novelist Jo Nesbø (2009). First Kahneman,

> A team of firefighters entered a house in which the kitchen was on fire. Soon after they started hosing down the kitchen, the commander heard himself shout "Let's get out of here!" without realizing why. The floor collapsed almost immediately after the firefighters escaped. Only after the fact did the commander realize that the fire had been unusually quiet and that his ears had been unusually hot ... He had no idea what was wrong, but he knew something was wrong. (Kahneman, 2011, 11)

Now Nesbø. In the story, Harry is the hero, Harry Hole, and Beate is a talented forensic detective.

> 'Forget what you have or haven't got,' Harry said. 'What was your first impression? Don't think, speak.'
>
> Beate smiled. She knew Harry now. First, intuition, then the facts. Because intuition provides facts too; it's all the information the crime scene gives you, but which the brain cannot articulate straight off. (Nesbø, 2009, 126)

There's at least a family resemblence between Harry Hole's instruction here and Lewis's instruction to his readers at the start of "Elusive Knowledge" (Lewis, 1996b).

> If you are a contented fallibilist, I implore you to be honest, be naive, hear it afresh. 'He knows, yet he has not eliminated all possibilities of error.' Even if you've numbed your ears, doesn't this overt, explicit fallibilism still sound wrong? (Lewis, 1996b, 550)

Reviewers of Nesbø's books often describe his hero as 'intuitive'. That's a little misleading; Harry Hole thinks intuition has a key role to play in detective work, but the adjective suggests that he relies heavily on his own intuition. That's not right; he's just as often badgering his colleagues to give him their impressions of a crime scene, or an interview subject. In these scenes he reminds me of no one so much as a colleague constantly wanting to know what one thinks about some thought experiment or variation on a familiar case. (These are often the best kind of colleague - full of inspiring ideas!)

So I think a lot of philosophical progress is made by drawing on, and drawing out, these skills. But isn't this just to say something uncontroversial and uninteresting, namely that philosophy relies on implicit knowledge? As Cappelen puts it,

> It is not controversial that conversations have propositions in the common ground. Nor is it controversial that all arguments start with premises that are not argued for. (155)

Well, there's something a bit interesting here, namely that the 'common ground' and the 'not argued for' premises have much greater overlap in philosophy than in other fields. A book starting with observations about the Galápagos Islands starts with premises that are not argued for, but are asserted on the basis of observations. These premises surely weren't in the common ground before the 'conversation' starts. I'll say more about this in the next section.

Because first I want to fuss a little about just what 'common ground' is. We'll start with an observation Cappelen makes about the Ginet/Goldman case of Henry and the fake barns (Goldman, 1976). Many philosophers take it to be an interesting fact that in one scenario, Henry knows there's a barn, while in another he does not. Cappelen says that these facts are "presented as being pre-theoretically in the common ground" (172). That seems false at first blush. Before reading Goldman's paper, it's not clear philosophers are in a position to form singular thoughts about Henry. That's an uncharitable reading though. A more plausible claim is to say that we are pre-theoretically disposed to accept some long sentence that roughly says that an agent in such-and-such scenario knows there is a barn, while an agent in a slightly different scenario does not.

We might gloss that last claim as saying that we implicitly knew something about these scenarios. I'm not sure that's right though. We do surely have lots of implicit knowledge. I know, and so do you, that the Sydney Opera House is south of the Royal Albert Hall, even if you'd never articulated that thought to yourself or another. But do our dispositions to respond to quite finely drawn, and often reasonably long, vignettes count as implicit beliefs, or should they count as things we were in a position to know, but only learned once a philosopher had done the work of drawing the vignette? I can see merit in both positions, and don't see firm grounds for preferring one.

Let's introduce some terminology to avoid taking a stance on this question. Say that a subject has *Socratic knowledge* that p when the following two conditions are met:

1. Once the agent is asked to consider p in the right way, they will come to know p.
2. The evidential basis for this knowledge that p is not the asking itself.

The first clause says that anyone who reacts to a Gettier case with "Oh, of course that's justified true belief without knowledge" has Socratic knowledge that such a case is a counterexample to the JTB theory of knowledge. And they have this Socratic knowledge before the case is even raised. The second clause says that if the person reacts instead with "Oh, some philosophers use thought experiments that don't make sense unless you know which cars come from which countries", that *won't* count as Socratic knowledge. They would be expressing some knowledge, to be sure, but the telling of the example would play an evidential role.

If you are very liberal about which dispositions count as implicit beliefs, and implicit knowledge, then Socratic knowledge will just be a special kind of implicit knowledge. But if you think considering examples can lead to learning new facts, not just drawing out dispositions, then you will think 'Socratic' is like 'alleged', a non-factive modifier. As I've defined it, once you hear Gettier cases once, that they are counterexamples to the JTB theory ceases to be Socratic knowledge, and becomes regular knowledge. Note also that we can make sense of some implicit states being more or less Socratic than others; some dispositions to assent require very careful work to trigger.

Why is the class of propositions that we Socratically know so rich and fertile? It's because of the central role of heuristics in our cognitive lives. Our interactions with the world don't just furnish us with a set of truths about the world. They also furnish us with skills that we can apply to generate more truths. I suspect that something like this observation is at the heart of the endorsement of F3, that intuitions reveal conceptual truths. When we intuit that p, we don't always merely recall a prior belief that p, or infer p from what we antecedently explicitly knew. But nor do we observe that p. So what is it? It must be something internal, but not memory or inference. Conceptual competence isn't a bad first guess, but Cappelen shows that isn't the right answer. I think the right answer has to do with cognitive skills, i.e., heuristics.

4 Philosophy: A Negative Characterisation

So intuitions matter because they reveal Socratic knowledge, and Socratic knowledge, when made explicit, is a very good guide to the world. That implies that intuitions should not be confined to philosophy. And, indeed, they are not. If an economic theorist claimed the standard of living among English men was higher in 1915 than in 1935, it would be perfectly reasonable to reply that intuitively that cannot be right, because in 1915 a rather large number of English men were living on the Western Front in catastrophically poor conditions. What is distinctive of philosophy then?

We need to clarify this question before we can answer it. Philosophy is both a discipline with a history over many millennia, and an organisational unit inside modern universities. These two things overlap well, but not perfectly. Once we note that they are distinct, we can separate out the following three questions.

1. What questions are philosophical questions?
2. What questions are, within the academy, primarily addressed by researchers in philosophy departments?
3. What questions should be, at least within the academy, primarily addressed by researchers in philosophy departments?

The three questions don't overlap. When Milton Friedman (1953) writes about economic methodology, I think he's addressing a philosophical question, but work like this is, and probably should be, carried out in economics departments. Questions about professional ethics are philosophical questions that I think should be researched in philosophy departments, but in the United States at least typically receive more attention in professional schools. Let's focus on the third question; what should a philosophy department do?

My colleagues at Michigan and St Andrews work on an incredibly wide range of questions, from the interpretation of quantum physics through history of logic through moral psychology and so on. And I think philosophy departments should have this range of interests. But what do all these questions have in common?

It's not anything to do with necessity or a priority. Those categories seriously cross cut philosophy, as Cappelen points out. Historical investigations into disputes about the parentage of various might-have-been-royals, or mathematical investigation into the nature of the primes are not philosophical, but have to do with necessity and a priority. Whether there's a language of thought is contingent, a posteriori, and almost paradigmatically philosophical.

It's not really anything to do with *depth*, at least on a natural understanding of that. Why pandas have thumbs, and humans have appendices, turn out to be reasonably deep questions, but they are for biologists, not philosophers. Under what circumstances is democracy compatible with a strong executive is, at least to me, an incredibly deep and important question, but it's a question to be answered, primarily, in history and political science departments.[2] On the other hand, whether we can tell a plausible supervaluationist story about belief reports is not particularly deep, but a perfectly good subject for a philosophical inquiry as in Weatherson (2003a).

Better, I think, is to say that philosophical questions are those where implicit or Socratic knowledge, including crucially intuitions, can plausibly play a large role in getting to an answer. Philosophy is a little recursive, so it includes investigations into its own investigations, including historical work and metaphilosophical work. (Two fields which, prior to Cappelen's book, had surprisingly little interaction.) That's not to say we're always right that Socratic knowledge can answer the questions philosophy sets. Maybe some questions in mind and language are best answered with the aid of neurological or phonological work that requires powerful measuring devices. But the questions are ones where starting with the knowledge and skills we already have seems like a plausible starting point, or at least not entirely crazy. This makes philosophy distinct from, say, history. We use intuitions in history too, especially intuitions about what explains what. But we need more; intuition won't help if you want to know how many troops Henry had at Agincourt.

This hypothesis explains, I think, one of the historically important facts about philosophy. Philosophy gives birth to disciplines. Physics, economics, psychology and cognitive science were all, at one time, part of philosophy. In some cases, the split was very recent. The economics tripos at Cambridge only split from philosophy in 1903 (Tribe, 2002). The *Australasian Journal of Philosophy* was the *Australasian Journal of Psychology and Philosophy* until 1946. Why does philosophy give rise to disciplines like these?

[2]This is not to say that political philosophers couldn't help with this question. There are lots of questions that should have as their research centre some other department, but to which philosophers can usefully help. Indeed, the examples from economic methodology and evolutionary explanation I just mentioned are two more such questions.

I think having a negative characterisation of philosophy helps explain it. Philosophy has a lot in common, methodologically, with physics, economics, psychology and so on. All those fields use intuitions and other forms of Socratic knowledge. But the other fields use other things too, especially observation. It's when it becomes clear that armchair methods play too small a role in the research that the field leaves philosophy.

Of course, philosophers care more about their questions than their methods, so when the need for non-armchair methods becomes pressing, some of the individual philosophers will go along, picking up more and more observational knowledge and experimental skills. Note how much more empirical research informs the recent work by (for example) Gilbert Harman, Kim Sterelny and Peter Carruthers, compared to their earlier work (Harman, 1973; Kilkarni and Harman, 2011; Devitt and Sterelny, 1987; Sterelny, 2012; Carruthers, 1990, 2011). From the other direction, our armchairs come with more knowledge now than they used to, which is partially why engaging with Laura Ruetsche's work in philosophy of science requires more empirical knowledge engaging with William Whewell's (Ruetsche, 2011; Whewell, 1840). But still I think the general picture holds; a question is fit for philosophy iff it is plausible that the intuitive, armchair methods which are part of every academic's toolkit can, on their own, generate serious progress on the question.

5 Letting Go

I've said that Lewis's instruction at the start of "Elusive Knowledge" is to look to intuitions, not to theoretical beliefs. But that might involve reading more into Lewis than is really there. What he literally asks the reader is to not appeal to their preferred theory of knowledge. Is that the same as an appeal to intuitions?

It need not always be. Sometimes, asking people to let go of their prior theory involves asking them to engage in a complex cognitive task. In Meditation One, Descartes has us go through quite a lot of thoughts before we can be pre-theoretical in the way he wants us to be.

But I don't think that's what's going on with Lewis. For one thing, he doesn't guide us back to a pre-theoretic naïveté the way Descartes does. But more generally, I think getting snap judgments is a way of letting go of some prior theories.

The picture I have here, and it is nothing more than a picture, is that intuitions are judgments delivered by heuristics, heuristics are deployed by Fodorian modules, and Fodorian modules are informationally encapsulated (Fodor, 1983, 2000). That is, when we rely on a heuristic, we don't use all of the information at our disposal. The classic example of this is eyesight; we may know that there are no elephants on Market Street in St Andrews, but given the right visual stimuli, our eyes will still insist that there is an elephant *right there*. The background theory about the spatial distribution of elephants isn't encoded into the visual module. More generally, to rely on a heuristic just is to make a judgment using a part of our mind that doesn't believe some of the things that we do. And that's good, because it is a kind of independent check on the beliefs we have.[3]

But isn't the idea that snap judgments are essential to philosophy inconsistent with the fact that we work very hard on getting our examples just right, and (as Cappelen shows), argue at

[3]Philosophers sometimes understate the importance of independent checks. We can know a scale is working, but if we want to check its reliability we don't use it, we use something else. I suspect that a certain amount of theory-independence is part of the explanation of the value of intuitions.

great length over what to say about various examples? I think it isn't, because there are two respects in which our practice reveals a sensitivity to snap judgments, and a respect for their use as a check on theorising.

Let me tell you a small secret. I haven't heard anything that even sounds like a counterexample to the broadly Stalnakerian theory of indicative conditionals that I like for about a decade. That's not because there aren't any intuitive counterexamples. It's just because my intuitions have been trained to accord with this kind of theory.[4] So what do I do? Do I give up on the use of intuitions as a test of theory? No, I ask colleagues for their intuitions. Sometimes I ask them a lot of different questions, and sometimes I work rather hard on refining the question, or (when they sadly disagree with my theory) finding ways to undermine their intuitions. Given the number of similar questions I get from other colleagues, I don't think my methodology here is distinctive. In short, we can work very hard before and after getting the snap judgments, while giving those judgments a role.

This might be more idiosyncratic, but I also do a bunch of things in papers to draw out snap judgments. The main idea is to distract the reader from the fact that they are about to be prompted for an intuition, one that may not accord with their preferred theory. So I'll use deliberately absurd props (like Vinny the talking vulture), or start an example without flagging that it is an example. My favourite move along these lines is to set up an example in such a way that the example doesn't make sense unless some theoretical claim I want to argue for is true. Then, after much discussion of the correct verdict on the case, I can announce that the very sensibility of the prior discussion is proof that, at least intuitively, the theory I'm pushing must be true.

We're going to come back to this theme a bit later, because I think it's rather important. The cases you can remember from papers are probably not the ones where intuition mattered. The big role for intuition in philosophy (and in many other disciplines) is in checking the small steps along the way. That's why I join Cappelen in opposing the methodological rationalists; I don't think intuitions are distinctive to philosophy, and these small steps don't have much of a phenomenology. But that doesn't mean they are unimportant.

6 Strength and Fragility

One of the big trends in late 20th Century epistemology has been the separation of two senses of *strength of evidence*. This might mean

1. How strong a doxastic state is supported by the evidence.
2. How resilient the force of the evidence is in the face of counterevidence.

One thing that conservative epistemologies (e.g., Harman (1986)) and dogmatic epistemologies (e.g., Pryor (2000)) have in common is that sources which might be very strong in the first sense might be very weak in the second sense. In particular, there can be sources of evidence that ground knowledge, and hence be rather strong in tthe first sense, but easily overturned by conflicting evidence. I prefer to reserve the terms 'strong' and 'weak' for the first sense, and use the terms 'resilient' or 'fragile' for the presence or absence of the second property. In that

[4] Relatedly, I haven't seen Liverpool get awarded an undeserved free kick for about that long.

language, the important insight of the conservatives and dogmatists is that evidence can be strong but fragile.

That's roughly how I think of intuitions – they are strong but rather fragile. So they can be unjustified justifiers, which is how I read Cappelen's feature F2 (i.e., Rock).[5]

Cappelen notes it is hard to tell whether something is being used as a starting point, or an unjustified justifier, so he gives three diagnostics for this. I mostly agree with one, and disagree with the other two. I agree that intuitions are non-inferential, and they aren't based on any particular experience, which is his criteria F2.1. (Though they usually are based on experiences taken collectively.) But I would alter the following suggestion, which he gives as a second diagnostic.

> F2.2 **Evidence Recalcitrance**: Intuitions are evidence recalcitrant; i.e., holders of them are not disposed to give them up even when their best arguments for those intuitions are shown to fail. (Compare pg 112)

I would rather offer something normative here. What's true of intuitions is that they might provide a stronger ground for belief than the best evidence we can offer for them. Compare the case of Gettier. As Cappelen carefully notes (194n3), Gettier doesn't appeal to a raw intuition. He gives an argument that his subjects don't know. Unfortunately, it isn't a compelling argument, since it takes as a premise that we can't get knowledge from a false belief, and that isn't quite right (Warfield, 2005). But Gettier was, to some extent, justified in believing these subjects didn't know to a greater degree than he was justified in believing this argument was sound. And that, I think, is not uncommon.

This is why I don't think Cappelen's 'Rough Guide to Rock Detection' (121), the third of the diagnostics, is perfectly reliable. He says that if evidence is given for p in a context, that's evidence that p isn't an unjustified justifier in that context. But sometimes we give arguments for judgments that we think could rest without them. Compare this little dialogue.

> A: Is 'John happiness' a well-formed sentence?
>
> B: No; it doesn't have a verb.

Here B gives a judgment, then offers a little argument for it. The argument has a strong premise, namely that all sentences have verbs. That's debatable; 'Lo, gavagai!' may be a counterexample. But B's judgment isn't undermined by examples that undermine her argument. As in the

[5] There's an ambiguity in Cappelen's text that I'm not sure I'm interpreting the right way. Let's that someone intuits that in a particular case, c doesn't cause e. Call the content of that intuition, i.e., what is *intuited*, p_d. And call the proposition that the person has this intuition, i.e., the event of the intuiting, p_g. Plausibly both p_d and p_g could be evidence in the right cases, though most of the time the salient evidence will be p_d. I think p_d can be an unjustified justifier in the sense that other beliefs, e.g., that a particular theory of causation is false, can be justified on the basis of p_d, but no other beliefs the agent has justify p_d. But you might want a stronger sense of 'unjustified', where it means not just not justified by anything else, but not justified *at all*. I think in these kinds of cases, p_d is justified, just not justified by anything else. And the justification is, as I'll get to below, strong but fragile. If when Cappelen says that intuitions, according to Centrality, are unjustified justifiers he means that the belief that p_d is unjustified, then I'm not defending Centrality. I just mean that the agent need not have any other mental states which justify the belief p_d, or indeed any access to anything that justifies p_d. But for all that it might be that the belief that p_d is justified, and the grounds for the justification include what the agent learned about causation as a child, plus perhaps her competence in distinguishing causes from non-causes.

Gettier case, we may give an argument that doesn't capture the full normative force of the judgment.

To say that intuitions are unjustified justifiers is not to say they are particularly special. If some conservative or dogmatic epistemology is true, there will be other unjustified justifiers. And if not, then this story about intuitions will be pretty implausible.

This picture of intuitions as strong but fragile meshes well, I think, with the picture from section 5. There I said the important intuitions are the ones you barely notice or remember. That's because the intuitions are fragile; if you remembered them enough to argue about them (or experimentally test them), the fragility conditions had probably been triggered, and the intuition probably wasn't doing much argumentative work.[6]

But why not think that intuitions are so fragile that they have no use in any philosophical debate? This question deserves more space than I can give it, but here are three sketches of answers.

1. Intuitions might be valuable checks on theory, and might be resilient enough to perform a valuable checking role.
2. Just like heuristics have characteristic errors, it might be that careful reasoning has characteristic errors, and there are cases where our first impressions are more reliable. See Gladwell (2005) for a summary of some relevant evidence.
3. Somewhat surprisingly, there may be cases when it is best to trust the less reliable source. The case for this is a bit detailed, and not original to me, so I'll just include a brief footnote for those interested.[7]

7 Some Lewisian Case Studies

I've described one kind of mental state that deserves the name 'intuition', and which could play a role in philosophical activity. But, as Cappelen presses, we have to work to convert that 'could' to a 'does'. Do we really rely in intuitive, or heuristic-driven, judgments about cases in analytic philosophy?

As Cappelen shows, the answer is "A lot less than you may have guessed." We argue a lot more than we intuit, especially about the famous cases.[8] The bit of analytic philosophy

[6] I'm simplifying a little here. My preferred position is that intuit*ed*s provide strong but fragile evidence, while intuit*ing*s provide weak but resilient evidence. The reason this is relevant is related to footnote 7.

[7] At one point in Ben Levinstein's doctoral dissertation, he considers whether there's a general rule for deciding which of two conflicting sources we can trust. There turns out to be very little in general one can say. In particular, *trust the more reliable source* turns out not in general to be good advice. If sources have characteristic errors, it might be that given what the two sources have said, it is better on this occasion to trust the less reliable source, because the verdicts the sources deliver provide evidence that we are seeing one of the characteristic errors of the more reliable source. It takes more space than I have here to fill in the details of this argument, and most of the details I'd include would be Levinstein's not mine. But here's the big conclusion. Assume that intuitions are often wrong, but rarely dramatically wrong. The reason for that is that heuristics are bad at getting things exactly right, and good at getting in the ballpark. And that careful reasoning is often right, but sometimes dramatically wrong. This is trickier to motivate, but I think true. Then when intuition dramatically diverges from theory, and we don't have independent reason to think that intuition is mistaken about the kind of case that's in question, we should trust the intuition more than the theory.

[8] There is interesting work to be done on the relative role of intuitions and arguments about *principles*, but I'm going to leave that for another day, and focus here on cases. The principles/cases distinction can be a bit slippery, but paradigm cases are easy to identify, and we'll be working with fairly paradigmatic cases here.

I'm most familiar with is David Lewis's corpus, and since that doesn't play much of a role in Cappelen's story, I'll illustrate his point with some examples from it.

Going from memory, I would have guessed the clearest example of a case refuting a theory was the use of finkish dispositions to refute the conditional analysis of dispositions. But go to the opening pages of "Finkish Dispositions" (Lewis, 1997b), and you find not an intuition about a case, but an argument that finks are possible. And even though that argument is followed up with more cases, Lewis rather explicitly *argues* for his conclusions about each one. See, for example, the glass loving sorcerer on page 147. Lewis doesn't avert to an intuition that the loved glass is fragile, rather he "wield[s] an assumption that dispositions are an intrinsic matter." (Lewis, 1997b, 147)

The discussion of causation turns out to be a little more fertile. From (the longer version of) "Causation as Influence", I count the following appeals to intuitions about cases.

- The chancy bomb example which shows simple probabilistic analyses of indeterministic causation won't work (Lewis, 2004a, 79).
- The Merlin and Morgana example which shows that trumping is possible, and matters for what is the cause (Lewis, 2004a, 81).
- The variant on Billy and Suzy that raises problems for quasi-dependence (Lewis, 2004a, 83).
- The crazed President example which shows that causation by double prevention is possible, and that causation is not an intrinsic relation (Lewis, 2004a, 84).
- The Frankfurt example which shows we can have causation without dependence (Lewis, 2004a, 95).

There's a strong sense, I think, in which none of the judgments in these cases are argued for. Indeed, they arise as *problems* for theories that are otherwise doing rather well. If there was an argument around, it would be for the negation of the intuited judgment. So I think there's a role for intuition here.

But we should not imagine that this is normal for philosophers, or even for Lewis. Cases, it is true, play a large role in Lewis's writing. But they are very rarely simple refutations of existing theories. We could perhaps distinguish four roles that cases play, or perhaps four types of philosophical cases.

1. Refutation of theories, as in these causation cases.
2. Illustrations that help explain what's going on in an argument, as in the examples from "Finkish Dispositions". For a more extensive version of this, see Lewis's version of Puzzling Pierre (Lewis, 1981c).
3. Tools for showing that we must distinguish various concepts, such as the discussion of Ned Kelly's proof that there's no honest cop (Lewis, 1988c).
4. Simplified versions of the real world, on which we can test various explanatory hypotheses, such as the footy and rugby people in "Naming the Colors" (Lewis, 1997c).

And that list is probably incomplete. The last is fairly fascinating as a case study actually.[9] Some of you may have had the following experience when programming, or indeed doing anything

[9] See Sugden (2000, 2009) for much more on this use of thought experiments.

that looks like working with code (such as writing in LaTeX). A bug arises. It helps to find a minimal example in which the bug arises, i.e., a smallest program that produces the same bug. This helps you spot what's going on, and if you still need help, it helps your interlocutors focus on the central problem. It's important that you haven't changed the problem; the example must be of the same kind as what you started with. But the example could be much simpler than the case you're most interested in. Some philosophy examples are, I suspect, like that. Their value lies in revealing that some striking feature of reality would persist even if the world were simpler. So, probably, the explanation of the feature lies in some respect the real messy world shares with the simple example world. (Compare Cappelen's discussion of Perry's messy shopper in section 8.1.)

It is perhaps no coincidence that the easiest place to find examples of type 1 in Lewis's work is in the papers on causation. Lewis thinks there is no such thing as causation (Lewis, 2004a,c). Whatever our theory of 'causes' should be, it shouldn't match that verb with a binary property. Rather, the aim of philosophical work on causation is to give a reductive analysis of causal thought and talk. In such a project, judgments about how we use 'causes' are more likely to be central.

It's also not coincidental that when an example is central to a paper, such as the 'dishonest' cop and Puzzled Pierre, they really don't look like type 1. That's one big and important lesson from Cappelen's work. Philosophers do use examples to refute theories, but they are rarely the big famous examples. If an example is central to a philosophy paper, it typically plays one of the other three roles.

8 Summary

Let's take stock. I've argued for the following theses:

1. Socratic knowledge is important to philosophy.
2. The distinctive feature of philosophy is that it addresses questions that can, at least *prima facie*, be productively worked on while relying primarily on Socratic knowledge.
3. Intuitions are manifestations of cognitive skills, and much Socratic knowledge is constituted by the possession of such cognitive skills.
4. Like other forms of Socratic knowledge, intuitions are mostly *a posteriori*, and have roles outside philosophy as well as inside it.
5. Intuitions are default justified; that is, they can be unjustified justifiers.
6. This default is very weak; intuitions can easily be overridden by other considerations.
7. Relatedly, it is rare for any one intuition to be central to a philosophical work; philosophical intuitions mostly concern the little cases we see along the way to larger projects.

I also hinted at, without developing, an argument for

8. The right intuition can stop even a plausible theory dead in its tracks; and we have (thanks to Ben Levinstein) a mathematical model for why this can be so even if intuitions are much less reliable than theories.

I opened with a discussion of why it matters to philosophy's self-conception that point (4) is correct. Since Cappelen also endorses (4), I probably don't need to say more about that here. But I think there is more to say about (7).

The first thing I want to note is that (7) is of course consistent with Cappelen's textual research on important work in late analytic philosophy. In just about any thought experiment that you can remember, the intuitions about it don't carry much philosophical weight in the work in which it is introduced. The intuitions that matter are the little ones, the ones that go by so quickly that no one questions them and are largely forgotten by all but the *cognoscenti* in that field. Even these intuitions aren't *that* common. There are less of them in Lewis than I would have guessed.

Still, I disagree with Cappelen that philosophy is without these intuitions. And so I disagree that there's no role for double checking, experimentally if need be, whether these intuitions are really intuitive. If a well run survey showed that most people disagree with Lewis's judgment about, say, the chancy bombs example, I'd reconsider my views about probabilistic causation. But I'd be really surprised to see this.

The second thing to note is that while (7) is true, it's not the case that intuitions about one case are never central to a philosophical project. There is one big counterexample: the Gettier literature. Like Cappelen (194n3), I think this literature is incredibly unrepresentative of philosophy. And I think that's in part because it was methodologically flawed. I tried to make this point in an earlier paper (Weatherson, 2003b), but I didn't get it quite right. (What I should have said was more like what Elijah Chudnoff (2011) does say.) When we saw the Gettier example, this should have been an invitation to try and find out what feature of knowledge was driving the fact that the belief in the main examples didn't amount to knowledge. Gettier suggested it was inference from a false premise, but that doesn't quite work (Warfield, 2005). You might think it is insensitivity, but that doesn't quite work. At this point there should have been one of two paths taken - attempts to find some other explanation of the data, or a reconsideration of whether our initial judgment about the case was wrong. That's what the picture of philosophy sketched here would have predicted, and (this is the point I was trying but failing to make in the earlier paper) that's what reflection on our successes in other areas of philosophy would have recommended. But the first kind of project ended up intertwined with attempts to analyse knowledge, and stalled for decades. And the second project wasn't seriously undertaken, with some honorable exceptions such as Sartwell (1992) and Hetherington (2001). Now eventually this didn't matter, because we discovered that safety based explanations of the Gettier case would work, even if there is no safety based analysis of knowledge, and even if there is some work to be done in getting the safety condition just right (Williamson, 1994, 2000; Sainsbury, 1995; Lewis, 1996b; Weatherson, 2004). So if we strengthened (7) into a universal claim it would be false – thirty years of epistemological struggle attest to this. But it was really when epistemology fell into line with practice in other fields of philosophy that it made progress on the Gettier case.

Keynes and Wittgenstein

Abstract
Three recent books have argued that Keynes's philosophy, like Wittgenstein's, underwent a radical foundational shift. It is argued that Keynes, like Wittgenstein, moved from an atomic Cartesian individualism to a more conventionalist, intersubjective philosophy. It is sometimes argued this was caused by Wittgenstein's concurrent conversion. Further, it is argued that recognising this shift is important for understanding Keynes's later economics. In this paper I argue that the evidence adduced for these theses is insubstantial, and other available evidence contradicts their claims.

1 Introduction

Three recent books (Davis, 1994; Bateman, 1996; Coates, 1996) have argued that the philosophy behind Keynes's later economics (in particular the *General Theory*) is closer to Wittgenstein's post Tractarian theorising than to his early philosophy as expressed in his *Treatise on Probability*.[1] If Keynes did follow Wittgenstein in the ways suggested it would represent a substantial change from his early neoplatonist epistemology. In this paper I argue that the evidence for this thesis is insubstantial, and the best explanation of the evidence is that Keynes's philosophical views remained substantially unchanged.

There are three reasons for being interested in this question. The first is that it is worthwhile getting the views of a thinker as important as Keynes right. The second is that it would be mildly unfortunate for those of us attracted to Keynes's epistemology to find out that it was eventually rejected by its creator[2]. Most importantly, all parties agree that Keynes thought his philosophical theories had substantial consequences for economic theory. It is a little unusual for philosophical theories to have practical consequences; if one is claimed to it is worthwhile identifying and evaluating the claim.

Section 2 examines Bateman's claim that Keynes abandoned the foundations of his early theory of probability. Bateman's arguments turn, it seems, on an equivocation between different meanings of 'Platonism'. On some interpretations the arguments are sound but don't show what Bateman wants, on all others they are unsound. Section 3 looks at the conventionalist, intersubjective theory of probability Bateman and Davis claim Keynes adopted after abandoning his early objective theory. As they express it the theory's coherence is dubious; I show how it might be made more plausible. Nevertheless, there is little to show that Keynes adopted it. The only time he talks about conventions is in the context of speculative markets and in these contexts a conventionalist theory will give the same results as an objectivist theory.

Section 4 looks at Coates's quite different arguments for an influence from Wittgenstein to Keynes. Part of the problem with Coates's argument is that the textual evidence he presents

† Unpublished.

[1] Davis's views are also set out in his (1995), and Coates's to some extent in his (1997), but I will focus on the more detailed position in their respective books.

[2] In the way that, for example, subjective Bayesianism was arguably invented by and eventually rejected by Ramsey. See his Ramsey (1926) and Ramsey (1929).

is capable of several readings; indeed competing interpretations of the pages he uses exist. A bigger problem is that even when he has shown a change in Keynes's views occurred, he immediately infers the change was at the foundations of Keynes's beliefs. Section 5 notes one rather important point of Wittgenstein's of which Keynes seemed to take no notice, leading to an error in the *General Theory*. This should cast doubt on the claim that Keynes's later philosophy, indeed later economics, was based on theories of Wittgenstein.

2 Bateman's Case for Change

A brief biographical sketch of Keynes is in order to frame the following discussions, though I expect most readers are familiar with the broad outlines[3]. Keynes arrived as an undergraduate at Cambridge in 1902 and was based there for the rest of his life. For the next six years he largely studied philosophy under the influence of Moore and Russell. In 1907 he (unsuccessfully) submitted his theory of probability as a fellowship dissertation; this was successfully resubmitted the following year. His plans to make a book of this were interrupted by work on Indian finance, the war and its aftermath. It appeared as *Treatise on Probability* (hereafter, *TP*) in 1921, after substantial work on it in 1920. Modern subjectivist theories of probability, generally known as Bayesian theories, first appeared in critical reviews of this book (e.g. Borel (1924), Ramsey (1926)). After leaving philosophy for many years, Wittgenstein returned to Cambridge in 1929, and subsequently had many discussions with Keynes. In Keynes's *General Theory* (hereafter, *GT*) of 1936 and in some of the ensuing debate, Keynes referred to some distinctive elements of the *TP*, leading some interpreters to suspect that there was a theoretical link between his early philosophy and his later economics.

There are two distinctive elements of Keynes's early theory of probability for our purposes. The first is its objectivism. Keynes held the probability of p given h is the degree of *reasonable* belief in p on evidence h, or, as Carnap (1950) put it, the degree of confirmation of p by h. These degrees are determined by logic; Keynes held that there was a partial entailment relation between p and h, of which the ordinary entailment relation (then thought to have been given its best exposition by Russell and Whitehead) was just a limiting case. And these relations are Platonic entities, we discover what they are by perceiving them through our powers of intuition. The second element is that the degrees may be non-numerical. So if the probability of p given h is α, we may be able to say $\alpha > 0.3$, and $\alpha < 0.5$, but not be able to give any finer numerical limits. As a corollary, there are now two dimensions of confirmatory support. Keynes claimed that as well as determining the probability of p given h, we could determine the 'weight' of this probability, where weight measures how much evidence we have. The more evidence is in h, the greater the weight. Keynes thought the distinction between saying that on evidence h, p has a low probability, and saying that the weight of that probability is low is important for understanding investment behaviour (*GT*: Ch. 12).

Bateman and Davis both claim that Keynes gave up this theory for an intersubjective theory in the *GT*. I'll focus on Bateman's book, largely because the structure of his argument is more straightforward[4]. Bateman sets himself to offer another solution to 'das Maynard Keynes

[3]For more details see Skidelsky (1983, 1992) or Moggridge (1992).

[4]All page references in sections 2 and 3 (unless otherwise stated) to Bateman 1996. Space considerations preclude a detailed examination of Davis's arguments, which are quite different to Bateman's. However his conclusions are subject to the same criticisms I make of Bateman's in section 3, and of Coates's in section 5.

problem', which he describes as follows.

> "[Future theorists] will read *Treatise on Probability*'s account of the objective nature of probabilities and the way that rational people employ them, and they will wonder at how this person could have turned around 15 years later and written a book [the *GT*] in which irrational people who base their decisions on social conventions cause mass unemployment in the capitalist system" (7)

I doubt this is the right thing to say about the *GT*, but that's another story. For now we might simply note that there's no obvious conflict here. For one thing, if the people in the *TP* are rational, and in the *GT* are irrational, as Bateman allows, it's not too surprising they behave differently. More generally, it's to be expected (sadly) that normative and descriptive theories are different, and by Bateman's lights that should explain the difference between the outlook of the explicitly normative *TP* and the at least partially descriptive *GT*. If the agents in the *GT* are irrational, that book cannot but be a purely normative account of rationality. On the other hand, if the *TP* were taken to be descriptive and not just normative, if it claimed that people really conform to its epistemological exhortations, there could be a conflict. I can't imagine, however, what the evidence or motivation for that reading could be.

If there were a conflict between the *GT* and *TP*, there ought be a greater one between the 'rational people' of the *TP* and the blatantly irrational leaders in *Economic Consequences of the Peace* (*ECP*). These books were published about 15 months apart, not 15 years. And the most memorable parts of *ECP* are the descriptions of the mental failings of President Wilson, who Lloyd George could 'bamboozle' into believing it was just to crush Germany completely, but not 'de-bamboozle' out of this view when it became necessary. Or maybe we should say there's a conflict because the characters in David Hume's histories do not meet his ethical or epistemological norms.

If we give up Bateman's claim that the actors in the *GT* are irrational, and substitute the claim that the norms of rationality in the two books differ, then we have a real conflict. And the most charitable interpretation of Bateman is that this is the conflict he intends to discuss. At the bottom of page 12 he goes close to saying exactly this, but then proceeds to support his position with evidence that Keynes changed his position on how rational people actually are. Once we are claiming the change of view is with regard to norms, evidence of opinion changes about empirical questions becomes irrelevant. This does mean much of Bateman's case goes, though not yet all of it.

The main problem with Bateman's argument is that it rests on an equivocation over the use of the term 'Platonism'. In *TP* Keynes held that probability relations are objective, non-natural and part of logic. I'll use 'logical' for the last property. When Bateman says Keynes believed probability relations were Platonic entities, he is alternately referring to each of these properties. He seems to explicitly use 'Platonic' to mean 'objective' on page 30, 'non-natural' on page 131, and 'logical' on page 123. But this isn't the important equivocation.

Say a theory about some entities is a 'Strong Platonist' theory if it concords with all Keynes's early beliefs: those entities are objective, non-natural and logical. Bateman wants to conclude that by the time of the *GT*, Keynes no longer had an objectivist theory of probability. But showing he no longer held a Strong Platonist view won't get that conclusion, because there are 3 interesting objectivist positions which are not Strong Platonist. The following names are my own, but they should be helpful.

Carnapian Probability relations are objective, natural and logical. This is what Carnap held in his 1950.

Gödelism Probability relations are objective, non-natural and non-logical. Gödel held this view about numbers, hence the name. I'd normally call this position Platonism, but that name's under dispute. Indeed I suspect this is what Keynes means by Platonism in *My Early Beliefs*. (Keynes, 1938b)

Reductionism Probability relations are objective, natural and non-logical. Such positions don't have to reduce probability to something else, but they usually will. Russell held such a position in his 1948.

These categories could apply to other entities, like numbers or moral properties or colours, but we will be focussing on probability relations here. Say a theory is 'Weak Platonist' if it is Strong Platonist or one of these three types. The most interesting equivocation in Bateman is using 'Platonist' to refer to either Strong or Weak Platonist positions. He argues that Keynes gave up his early Platonist position. These arguments are sound if he means Strong Platonist, unsound if he means Weak Platonist. But if he means Strong Platonist he can't draw the extra conclusion that Keynes gave up objectivism about probability relations, which he does in fact draw. So I'll examine his arguments under the assumption that he means to show Keynes gave up Weak Platonism.

Whatever Bateman means by Keynes's Platonism, he isn't very sympathetic to it. It gets described as 'obviously flawed' (4) and 'fatally flawed' (17), and is given as the reason for his work being ignored by 'early positivists and members of the Vienna Circle' (61). Given that the *TP* is cited extensively, and often approvingly, by Carnap in his 1950, this last claim is clearly false. Most stunningly, he claims writers committed to the existence of Platonic entities cannot 'be considered to be a part of the analytic tradition' (39), though he does concede in a footnote that some 'early analytical philosophers' (he gives Frege as an example) were Platonist. Bateman's paradigm of philosophy seems to be the logical positivism of Ayer's *Language, Truth and Logic*: "nowhere would one less expect to find metaphysics than in modern analytical philosophy" (Ayer, 1936, 39).

There is an implicit argument in this derision. Keynes must, so the argument goes, have given up (Weak) Platonism because no sensible person could believe it. If anything like this were sound it should apply to Weak Platonism about other entities. But the history of 'modern analytical philosophy' shows that Weak Platonism (though not under that name) is quite widespread in metaphysical circles. Modern philosophy includes believers in possible worlds both concrete and ersatz, in universals and in numbers. All these positions would fall under Weak Platonism. Even Quine's ontologically sparse *Word and Object* was Weak Platonist about classes, though he probably wouldn't like the label. So by analogy Weak Platonism about probability relations isn't so absurd as to assume Keynes must have seen its flaws.

Bateman's more important argument is direct quotation from Keynes. This argument is undermined largely because of Bateman's somewhat selective quotation. There are two sources where Keynes appears to recant some of his early beliefs. Which early beliefs, and how early these beliefs were, is up for debate. The two are his 1938 memoir *My Early Beliefs* (hereafter, *MEB*), and his 1931 review of Ramsey's posthumous *Foundations of Mathematics*. *MEB* wasn't published until 1949, three years after Keynes's death, but according to its introduction it

is unchanged from the version Keynes gave as a talk in 1938. In it he largely discusses the influence of Moore, and particularly *Principia Ethica*, on his beliefs before the first world war.

There are several connections between Moore's work and Keynes. The most pertinent here is that Keynes's metaphysics of probability in *TP* is borrowed almost completely from Moore's metaphysics of goodness. Not only are probability relations objective and non-natural, they are simple and unanalysable. These are all attributes Moore assigns to goodness. The only addition Keynes makes is that his probability relations are logical. So Moore's position on goodness is, in our language, Gödelian.

As he says in *MEB*, Keynes became convinced of Moore's metaethics, though he differed with Moore over the implications this had for ethics proper. In particular he disagreed with Moore's claim that individuals are morally bound to conform to social norms. Bateman seems to assume that at any time Keynes's metaphysics of goodness and probability will be roughly the same, and with the exception of questions about their logical status, this seems a safe enough assumption.

Bateman quotes Keynes saying that his, and his friends', belief in Moore's metaethics was 'a religion, some sort of relation of neo-platonism' (Keynes, 1938b, 438). This is part of the evidence that Keynes meant what I'm calling Gödelism by 'Platonism'. Not only does he use it to describe Moore's position, but comparing Platonism with religion would be quite apt if he intends it to involve a commitment to objective, non-natural entities. The important point to note is that he is using 'religion' to include his metaethics, a point Bateman also makes, though it probably also includes some broad ethical generalisations. Bateman then describes the following paragraph as removing 'any doubt that [Keynes] had thrown over his youthful Platonism as untenable'. (40)

> Thus we were brought up – with Plato's absorption in the good in itself, with a scholasticism which outdid St. Thomas, in calvinistic withdrawal from the pleasures and successes of Vanity Fair, and oppressed with all the sorrows of Werther. It did not pervert us from laughing most of the time and we enjoyed supreme self-confidence, superiority and contempt towards all the rest of the unconverted world. But it was hardly a state of mind which a grown-up person in his senses could sustain literally. (Keynes, 1938a, 442).

As it stands, *perhaps* the last sentence signals a change in metaphysical beliefs, as opposed to say a change in the importance of pleasure-seeking. In any case the following paragraph (which Bateman neglects to quote) shows such an interpretation to be mistaken.

> It seems to me looking back, that this religion of ours was a very good one to grow up under. It remains nearer the truth than any other I know, with less extraneous matter and nothing to be ashamed of ... It was a purer, sweeter air than Freud cum Marx. It is still my religion under the surface. (Keynes, 1938a, 442).

So was Keynes confessing to 'a state of mind which a grown-up person in his senses couldn't sustain literally'? No; his 'religion' which he held onto was a very broad, abstract doctrine. It needed supplementation with a even general ethical view, to wit an affirmative answer to one of Moore's 'open questions'. And then it needed some bridging principles to convert those ethics into moral conduct in the world as we find it. His early position included all these, and it

seems it was in effect his early 'bridging principles' he mocks in the above quote. These relied, the memoir makes clear throughout, on an excessively optimistic view of human nature, so he thought in effect that he could prevent wrong by simply proving to its perpetrators that they were wrong. Now giving up one's bridging principles doesn't entail abandonment of a general ethical view, let alone one's metaethics. Indeed, let alone one's metaphysics of probability! And as the last quote makes clear, Keynes was quite content with the most general, most abstract parts of his early belief. If this were all Bateman had to go on it wouldn't even show Keynes had abandoned Strong Platonism[5].

There is more to Bateman's case. In Keynes's review[6] of Ramsey (1931), he recanted on some of his theory of probability. This is quite important to the debate, so I'll quote the relevant section at some length.

> Ramsey argues as against the view which I had put forward, that probability is concerned not with objective relations between propositions but (in some sense) with degrees of belief, and he succeeds in showing that the calculus of probabilities simply amounts to a set of rules for ensuring that the system of degrees of belief which we hold shall be a *consistent* system. Thus the calculus of probability belongs to formal logic. But the basis of our degrees of belief – or the *a priori* probabilities, as they used to be called – is part of our human outfit, perhaps given us merely by natural selection, analogous to our perceptions and our memories rather than to formal logic. So far I yield to Ramsey – I think he is right. But in attempting to distinguish 'rational' degrees of belief from belief in general he was not yet, I think, quite successful. It is not getting to the bottom of the principle of induction to merely say it is a useful mental habit. (Keynes, 1931, 338-339).

Tellingly, Bateman neglects to quote the final two sentences. I think there is an ambiguity here, turning on the scope of the 'so far' in the fourth sentence. If it covers the whole section quoted, it does amount to a wholesale recantation of Keynes's theory, and this is Bateman's interpretation. But if we take the first sentence, or at least the first clause, as being outside its scope it does not. And there are two reasons for doing this. First, it seems inconsistent with Keynes's later reliance on the *TP* in parts of the *GT*, as (O'Donnell, 1989, Ch. 6) has stressed. Secondly, it is inconsistent with Keynes's complaint that on Ramsey's view induction is merely a 'useful habit'. If Keynes had become a full-scale subjectivist, he ought have realised that patterns of reasoning could only possibly be valid (if deductive) or useful (otherwise). Since he still thought there must be something more, he seems to believe an objectvist theory is correct, though by now he is probably quite unsure as to its precise form. So in effect what Keynes does in this paragraph is summarise Ramsey's view, list the details he agrees with (that probability relations aren't logical), notes his agreement with them, and then lists the details he disagrees with (that probability relations aren't objective).

There is more evidence that all this quote represents is a recantation of the view that probability relations are logical. Earlier in that review he notes how little formal logic is now believed to achieve compared with its promise at the start of the century.

---

[5]The above points are similar in all substantial respects to those made by O'Donnell (1991) in response to an earlier version of Bateman's account.

[6]This is often mistakenly referred to as an obituary in the literature, e.g. (Coates, 1996, 139).

> The first impression conveyed by the work of Russell was that the field of formal logic was enormously extended. The gradual perfection of the formal treatment at the hands of himself, of Wittgenstein and of Ramsey had been, however, gradually to empty it of content and to reduce it more and more to mere dry bones, until finally it seemed to exclude not only all experience, but most of the principles, usually reckoned logical, of reasonable thought. (Keynes, 1931, 338).

More speculatively, I suggest Keynes's change of mind here (for this shows he had surely given up the view that probability relations are logical) might be influenced by Gödel's incompleteness theorem. In the *TP* Keynes had followed Russell is saying mathematics is part of logic (Keynes, 1921, 293n). That view was often held to be threatened by Gödel's proof that there are mathematical truths which can't be proven, and that the consistency of mathematics can't be proven. But no one suggested this meant mathematics is merely subjective, or that mathematical Platonism was therefore untenable. If this response to Gödel is right, it shows there are objective standards of reasoning (i.e. mathematical standards) that are not part of logic. This makes it less of a leap to say there are objective principles of reasonable thought that are not 'logical' in the narrow sense we've been using.

So would Keynes have known of Gödel's theorem when he wrote this review? I think it's possible, though some more research is needed. Keynes's review was published in *The New Statesman and Nation* on October 3, 1931. This was a weekly political and literary magazine of which Keynes was chairman. So we can safely conclude the piece was drafted not long before publication. Gödel's theorem was first announced at a conference in Vienna in September 1930 (Wang, 1987), and was published in early 1931. While Keynes would certainly have not read Gödel's paper, its content could easily have reached him through Cambridge in that 12 month 'window'. Since the explicit aim of Gödel's paper was to show the incompleteness of *Principia Mathematica*, it would have immediately had some effect in Cambridge, both in philosophy and mathematics. Given this evidence, the probability Keynes knew of Gödel's theorem when he wrote the review of Ramsey still mightn't be greater than one-half, but it mightn't be less than that either.

In sum, I conclude that Keynes had given up his earlier belief that all rules of reasonable belief are logical. This is what he yields to Ramsey. This concession would be supported by the 'drying up' of formal logic that Keynes notes, perhaps most dramatically expressed in Gödel's theorem. But he hadn't given up the belief that there are objective rules which are extra-logical, and given the identification of probability with degree of reasonable belief, he had no reason to reject Gödelism or Reductionism about probability. Hence Bateman's argument that he rejected objectivist theories of probability fails.

3 Conventionalism

Bateman and Davis each argue that Keynes adopted a conventionalist, intersubjectivist theory of probability. In Davis this is explicity attributed to Wittgenstein's influence, however in Bateman it is less clear what the source of this idea is. It isn't obvious what they mean by an intersubjective theory. In particular, it isn't clear whether they mean this to be an empirical or a normative theory; whether Keynes is claiming that we ought set our degrees of belief by convention or that we in general do. Since the empirical theory would be consistent with his

objectivist norms, and they stress the change in his views, I conclude they are claiming this is a new normative view. According to this view being reasonable is analysed as conforming to conventions. This is not a very standard epistemological position, but something similar is often endorsed in ethics. Bateman marshals the evidence that Keynes moves from an objectivist to a conventionalist position in ethics as evidence for this epistemological shift, but this doesn't seem of overwhelming significance[7].

Here's the closest Bateman gets to a definition of what he means by an intersubjective theory of probability.

> When probabilities are formed according to group norms, they are referred to as intersubjective probabilities ... I take it to be the case that in a world of subjective probabilities some individuals will form their own estimates and others will form them on the basis of group norms (50n).

This makes it look very much like an empirical theory, as it refers to how people actually form beliefs, not how they ought. So his intersubjectivism looks perfectly consistent with Keynes's objectivism. I am completely baffled by the 'world of subjective probabilities'. I wonder what such a world looks like, and how it compares to our world of tables, chairs and stock markets?

Fortunately there is a theory that does the work Bateman needs. Ayer (1936) rejects orthodox subjectivism about probability on the grounds that it doesn't allow people to have mistaken probabilistic beliefs. But he can't admit Keynesian probability relations into his sparse ontology. The solution he adopts is to define probability as degree of rational belief, but with this caveat.

> Here we may repeat that the rationality of a belief is defined, not by reference to any absolute standard, but by reference to part of our own actual practice (Ayer, 1936, 101).

The 'our' is a bit ambiguous; interpreting it to refer to the community doesn't do violence to the text, though it is just as plausible that it refers to a particular agent. The 'part of our practice' referred to is just our general rules for belief formation. These aren't justified by an absolute standard; they are justified by the fact they are our rules, and presumably by their generality. Given Bateman's views about metaphysics, it seems quite reasonable to suppose he'd follow Ayer on this point.

The evidence Keynes adopted such a position is usually taken to be some passages from the *GT* and the 1937 *QJE* paper in which he replied to some attacks on that book. Here's the key points from the two quotes Bateman uses to support his view.

> In practice we have agreed to fall back on what is, in truth, a *convention*. The essence of this convention – though it does not, of course, work out quite so simply – lies in assuming that the existing state of affairs will continue indefinitely, except in so far as we have specific reasons for expecting a change (*GT*: 152).

[7] If Keynes had adopted a framework which implied a tight connection between epistemological and ethical norms, such as a form of utilitarianism that stressed maximisation of expected utility, this would be important, since he couldn't change ethics and keep his epistemology. But such frameworks aren't compulsory, and given the vehemence with which Keynes denounced utilitarianism (Keynes, 1938b, 445) it seems he didn't adopt one.

> How do we manage in such circumstances to behave in a manner which saves out faces as rational, economic men? We have devised for the purposes a variety of techniques, of which much the most important are the three following: ...
>
> (3) Knowing that our own individual judgement is worthless, we endeavour to fall back on the judgement of the rest of the world which is perhaps better informed. That is, we endeavour to conform with the behaviour of the majority or the average. The psychology of a society of individuals each of whom is endeavouring to copy the others leads to what we may strictly term a *conventional* judgement (Keynes, 1937, 115).

There are two problems with using this evidence the way Bateman does. The first is the old one that they seem expressly directed to empirical questions, though perhaps appearances are deceptive here. The more important one is that Keynes is attempting to answer a very specific question with these passages; in ignorance of the question we can easily misinterpret the answer.

How much ought one pay for a share in company X? Well, if one intends to hold the share come what may, all that matters is the expected prospective yield of X's shares, appropriately discounted, as compared to the potential yield of that money in other uses. But as Keynes repeatedly stresses (*GT*: 149; (Keynes, 1937, 114)) we have no basis for forming such expectations. Were this the only reason for investing then purely commercial investment may never happen.

There is another motivation for investment, one that avoids this problem. We might buy a share in X today on the hope that we will sell it next week (or next month or perhaps next year) for more than we paid. To judge whether such a purchase will be profitable, we need a theory about how the price next week will be determined. Presumably those buyers and sellers will be making much the same evaluations that we are. That is, they'll be thinking about how much other people think X is worth.

> We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practice the fourth, fifth and higher degrees (*GT*: 156).

There is simply no solution to this except to fall back on convention. That is, we are forced into a conventionalist theory of value, at least of investment goods. But this doesn't mean that we have a conventionalist epistemology. On the contrary, it means that our ordinary (objectivist) empiricism is unimpeded. For the question that Keynes has us solve by reference to convention is: What is the value of X? This is equivalent to, what will be value of X be, or again, to what are the conventional beliefs about X's value? We need to answer a question about the state of conventions, and as good empiricists we answer it by observing conventions.

An analogy may help here. Here's something that Hempel believed: to gain rational beliefs about the colour of ravens, one has to look at some birds. Did this mean he had an ornithological epistemology? No; he had an empiricist epistemology which when applied to a question about ravens issued the directive: Observe ravens! Similarly Keynes's belief that to answer questions about value, i.e. about conventions, one has to look at conventions, does not imply a conventionalist epistemology. It just means he has an empiricist epistemology which when applied to a question about conventions issues the directive: Observe conventions!

There might be another motivation for using conventions, again consistent with Keynes's objectivist empiricism. Sometimes we may have not made enough observations, or may not have the mental power to convert these to a theory. So we'll piggyback on someone else's observations or mental powers. (This seems to be what's going on in the quote from Keynes (1937).) Or even better, we'll piggyback on everyone's work, the conventions. To see how this is consistent with an objectivist epistemology (if it isn't already obvious) consider another analogy.

What is the best way to work out the derivative of a certain function? Unless your memory of high-school calculus is clear, the simplest solution will be to consult an authority. Let's assume for the sake of argument that the easiest authorities to consult are maths texts. It seems like the rational thing to do is to act as if the method advanced by the maths texts is the correct method. Does this mean that you have adopted some kind of authoritarian metaphysics of mathematics, where what it is for something to be correct is for it to be asserted by an authority? Not at all. It is assumed that what the textbook says is correct, but the authoritarian has to make the extra claim that the answer is correct *because* it is in the textbook. This is false; that answer is in the textbook because it is correct. In sum, the authoritarian gets the direction of fit wrong.

Similarly in the 'piggyback' cases the intersubjectivist gets the direction of fit wrong. We are accepting that p has emerged as 'average opinion', then it is reasonable to believe p. But we aren't saying with the intersubjectivist it is reasonable to believe p because p is average opinion; rather we are assuming p is average opinion because it is reasonable to believe p.

The evidence so far suggests Keynes's statements are consistent with his denying intersubjectivism. We might be able to go further and show they are inconsistent with his adopting that theory. After the quote on *GT* page 152 he spends the next page or so defending the use of conventions here. The defence is, in part, that decisions made in accord with conventions are reversible in the near future, so they won't lead to great loss. If he really were an intersubjectivist, the use of conventions would either not need defending, or could be defended by general philosophical principles. Secondly, there is this quote which in context seems inconsistent with adopting a conventionalist view.

> For it is not sensible to pay 25 for an investment which you believe the prospective yield to justify a value of 30, if you also believe that the market will value it at 20 three months hence (*GT*: 155).

The context is that he is discussing why reasonable professional investors base their valuations on convention rather than on long-term expectation. Hence the 'you' in the quote is assumed to be reasonable. Hence it is reasonable, Keynes thinks, to believe that an investment's prospective yield justifies a value of 30, and that conventional wisdom is that its prospective yield is much lower. But if all reasonable beliefs were formed by accordance with conventional wisdom, this would be inconsistent. Hence Keynes cannot have adopted a conventionalist epistemology.

4 Keynes and Vagueness

What a terrible state Keynes interpretation has got into! From the same few pages (the opening of *GT* Ch. 4) Coates (1996) reads into Keynes a preference for basing theory on vague predicates, Bradford and Harcourt (1997) read Keynes as denying that predicates which are unavoidably vague can be used in theory, and O'Donnell (1997) sees Keynes as holding a position in between these.

Coates's theory is that Keynes abandoned the narrowly analytic foundations of his early philosophy because of the problems of vagueness that were pointed out to him by Wittgenstein. He has Keynes in 1936 adopting a middle way between analytic and Continental philosophy, which gives up on analysis because of unavoidable vagueness, but which doesn't follow Derrida in saying all that's left after analysis is 'poetry'. He also wants to argue for the philosophical importance of this theory. In this essay I'll focus on his exegetical theories, though there are concerns to be raised about his philosophy.

As in Bateman, analytic philosophy gets very narrowly defined in Coates[8]. Here it includes the claim that truth-value gaps are not allowed (xii). This excludes from the canon some of the most important papers in analytical philosophy of the last few decades (e.g. Dummett (1959), van Fraassen (1966), Fine (1975b), Kripke (1975)), and hence must be a mistake. To use one of Coates's favourite terms, 'analytic philosophy' is a family resemblance concept, not to be so narrowly cast. In particular, as we'll see, analytic philosophers don't have to follow Frege in being nihilist about vagueness.

Even more bizarrely, Coates defines empiricism so it includes both psychologism in logic and utilitarianism in ethics (72-3). Since Ayer (1936) opposes each of these doctrines, does that makes Ayer an anti-empiricist? If Ayer is a paradigm empiricist (as seems plausible) Keynes's rejection of psychologism and utilitarianism can hardly count as proof of opposition to empiricism, as Coates wants it to do. Apart from the fact that Mill believed all three, there is no interesting connection between empiricism, psychologism and utilitarianism.

Coates's story is that in the *GT* Keynes allowed both his units and his definitions to be quantitatively vague so as to follow natural language. This constitutes a new 'philosophy of social science' (85) that is based on the ordinary language philosophy of the later Wittgenstein. There are several problems with this story. The first is that most of Coates's evidence comes from *obiter dicta* in early drafts of the *GT*; by the time the book was finished most of these suggestions are expunged. The second is that it's quite possible to accept vagueness within a highly analytic philosophical framework. The third is that the way Keynes uses vagueness is only consistent within such a framework.

The first part of the story focuses on how Keynes derided his predecessors for using concepts that were vague as if they were precise. Coates adduces evidence to show Keynes in this context used 'vague' as a synonym for 'quantitatively inexact'. The most important concept misused by Keynes's predecessors in this way was the general price level. Of course this was hardly a new point in the *GT*; Keynes (1909) says similar things. Coates claims that Keynes's reaction to this misuse was to 'criticise formal methods' (83), and to conclude that 'economic analysis can do without the "mock precision" of formal methods' (85). This is all hard to square with Keynes's explicit comments.

[8]All page references in this section (unless otherwise stated) to Coates (1996).

> The well-known, but unavoidable, element of vagueness which admittedly attends the concept of the general price-level makes this term very unsatisfactory for the purposes of a causal analysis, which ought to be exact (*GT*: 39).

Further, Keynes then defends his choice of units of quantity (quantity of money-value and quantities of employment) on the grounds that they are not quantitatively vague. Coates is surely right when he says that Keynes's analysis of vagueness here is 'not very controversial'; although it is perhaps misleading to say it is controversial at all.

The second, and central, part of the story focuses on how Keynes allowed his definitions to be vague, but defended this on the grounds of conformity to ordinary language. This 'introduces what is distinctive about his later philosophy of the social sciences' (85). The bulk of Coates's evidence comes from Keynes's commentary on his own definitions; usually this includes a claim that he has captured the ordinary usage of the term. Since he uses 'common usage' to explicitly mean 'usage amongst economists' (*GT*: 79) the support these *dicta* give to Coates's theory might be minimal, but we'll ignore that complication. The real problem is that this commentary extends to cases where he has changed his mind over the best definition. For example, Coates quotes Keynes writing in a draft of the *GT* about the definition of income.

> But finally I have come to the conclusion that the use of language, which is most convenient on a balance of considerations and involves the least departure from current usage, is to call the actual sale proceeds *income* and the present value of the expected sale proceeds *effective demand* (Keynes, 1934, 425).

Coates comments:

> By choosing definitions on the ground that they correspond with actual usage Keynes was formulating an ordinary language social science, one that bears a resemblance to those argued for by philosophers of hermeneutics (90).

He then goes on to note some comments from the *GT* apparently about this definition, and how it relates to common usage. The problem is that this isn't the definition of income Keynes settles on in the *GT*. There he defines income of an agent as "the excess of the value of his finished output sold during the period over the prime cost" (*GT*: 54), and *net income* (which Coates fails to distinguish) as income less supplementary cost. Given that at every stage Keynes justified his current definitions by their (alleged) conformity with common usage, even when he changed definitions, it is hard to believe that these justifications are more than rhetorical flourishes. After all, who will deny that *ceteris paribus* technical definitions should follow ordinary usage?

If Keynes's early choice of definitions showed an adherence to a 'philosophy of hermeneutics', perhaps his abandonment of those definitions constitutes abandonment of that philosophy. One change doesn't necessarily mean a change in foundations, so it is worth looking at those foundations.

As I mentioned, allowing that vagueness exists doesn't mean abandoning the Russellian program of giving a precise analysis of language. There are two reasons for this. First, contra Wittgenstein it is possible to analyse vague terms. Secondly, there are semantic programs very much in the spirit of Russell which allow vagueness. I'll deal with these in order.

In *Philosophical Investigations*, Wittgenstein argued that the existence of vagueness frustrated the program of analysis (ss. 60, 71). The argument presumably is that analyses are precise, and hence they cannot accurately capture vague terms. (See also his comments about the impossibility of drawing the boundaries of 'game' in s. 68.) This is a simple philosophical mistake. We can easily give an analysis of a vague term, we just have to make the analysans vague in exactly the same way as the analysandum.

To see this in action, consider that paradigm of modern philosophy, Lewis's analysis of subjunctive conditionals or counterfactuals. Lewis (1973b) says that the conditional 'If p were the case, it would be that q' is true iff q is true in the most similar possible world in which p. He considers the objection that 'most similar' is completely vague and imprecise.

> Imprecise it may be; but that is all to the good. Counterfactuals are imprecise too. Two imprecise concepts may be rigidly fastened to one another, swaying together rather than separately, and we can hope to be precise about their connection Lewis (1973b).

Whatever the fate of Lewis's theory, his methodology seems uncontestable. Wittgenstein's claim that analysis must be abandoned because of vagueness is refuted by these observations of Lewis. Hence Coates's claim that allowing vagueness (as Keynes does) means giving up on analytic philosophy is mistaken.

The second problem with Coates's comments on vagueness is that he hasn't allowed for what I'll call 'orthodox' responses to vagueness. The aim of the early analytics drifted between giving a precise model for natural language, and replacing natural language with an artificial precise language. The latter, claims Coates, ought be abandoned because of the pragmatic virtues of a vague language. Let's agree to that; can the spirit of the early aim of giving a precise analysis of language be preserved?

Two approaches which seem to meet this requirement are the supervaluational and epistemic theories of vagueness. The supervaluationist says language can't be represented by a precise classical model, but it can be represented by a set of such models. The epistemic theorist says that there is a precise model of language, but we cannot know what it is[9]. Call a theorist who adopts one of these approaches 'orthodox'. The name is chosen because supporters and critics of orthodoxy agree that these positions represent attempts to minimise deviations from the classical, Russellian program.

Clearly Keynes did not explicitly adopt an orthodox theory of vagueness. Williamson (1994) attempts to trace the epistemic theory back to the Stoics, but general consensus is that these approaches were all but unknown until recently. What I want to argue is that Keynes's intuitions are clearly with orthodoxy. Coates, on the other hand, wants to place Keynes in a tradition that is critical of classical analysis, and perhaps finds its best modern expression in the exponents of fuzzy logics. To see this is wrong, note that the following beliefs are all in the *GT*.

(1) All goods are (definitely) investment goods or consumption goods.
(2) For some goods it is vague whether they are an investment or consumption good. (*GT*: 61)

[9]See Williamson (1994) for the best epistemic account, Fine (1975b) and Keefe (2000) for the best supervaluationist accounts.

(3) The yield of an investment, q, is vague.
(4) The carrying cost of an investment, c, is vague.
(5) The net yield of an investment, q - c, can be precisely determined. (*GT*: 226)

Since Keynes believed (1) to (5) we can safely conclude he believed they were consistent. More importantly, since the *GT* has been analysed more thoroughly than any other economic text written this century, and no one has criticised the consistency of (1) to (5), it seems many people agree with him. Hence if conformity with pre-theoretic intuitions of consistency is a central desideratum of a theory of vagueness, we can discard any theory that does not say they are consistent. However, of those theories on the market, only orthodox theories meet this requirement. It might also be noted that (1) and (2) are repeated in just about every introductory macro textbook, again without to my knowledge any question of their consistency.

We can quickly see that these propositions are all consistent on either orthodox theory. The supervaluationist says there is a set of classical models for a language; a sentence is true iff it is true on all models, false iff it is false on all models, and truth-valueless otherwise. Vague terms have different meanings on different models. So for a particular good, say a car, about which it is vague whether it is an investment or consumption good, the supervaluationist says it is an investment good on some models and a consumption good on others. So (2) is satisfied; however on all models it, like everything else, is either a consumption or investment good, so (1) is satisfied. Similarly because it is vague whether some costs should be counted as deductions from the yield of an investment or increments to its carrying cost, the values of q and c will be different on different models. Hence (3) and (4) are true, but q - c is constant across models[10], so (5) is true.

The epistemic theorist says that vagueness is just ignorance. As we can know that a car is an investment or consumption good without knowing which, (1) and (2) can be satisfied. Similarly, since we can know that a cost is incurred without knowing how to account for it in Keynes's terms, we can know q - c precisely without knowing q or c precisely, and hence (3) to (5) can be satisfied.

The heterodox theorist has a harder time. The theorist who, following Russell (1923), says that vagueness is infectious, if a part is vague so is the whole, will deny that (1) and (2) can be true together. Unless it's definitely true that a car is an investment or definitely true it's a consumption good it can't be definitely true that it's one or the other. This also seems to be the position taken by Wittgenstein (1953).

The nihilist about vagueness, who follows Frege in saying vague terms can't be used coherently, similarly can't endorse both (1) and (2). On that view, if p and q are both vague, then their disjunction can't be true. Arguably, on this position the disjunction of p with anything can't be true, as it is nonsense, but we don't need anything that strong.[11]

The extra truth-values approach to vagueness (of which fuzzy logic is a variant) also can't make (1) and (2) consistent. On any such approach (whether 3-valued, n-valued or continuum-valued) the degree of truth of a disjunction can't be higher than the degree of truth of each of the disjuncts. So if neither 'This is an investment' nor 'This is a consumption good' is absolutely

[10] A particular cost will either remove an amount from q or add an equal amount to c, depending on how it is categorised.

[11] Compare the logic in Bochvar (1939), where $p \vee q$ is truth-valueless if p is true and q truth-valueless. Summaries of this and many other many-valued logics are in Haack (1974).

true (true to degree 1), 'This is an investment or consumption good' can't be absolutely true. Yet this is just what Keynes asserted to be possible, and what several generations of readers have found perfectly consistent. I have only remarked about the problem the consistency of (1) and (2) poses for heterodox theories. These remarks apply, *mutatis mutandis*, to (3), (4) and (5), but as theorists rarely discuss quantitative vagueness (as opposed to truth-value vagueness) these cases involve a bit more speculation as to what heterodoxy says.

Hence Keynes did not belong to a heterodox tradition *vis a vis* vagueness, and heterodox theories fail to capture a crucial pre-theoretic intuition about vague terms. So Coates's claims that Keynes followed Wittgenstein into heterodoxy here, and that he ought have, are both mistaken.

Even if all of the above is mistaken, there remains serious doubt that Keynes had in mind anything like what Coates attributes to him. Coates makes the chapters on definitions in the *GT* into the foundations of a new philosophy, and constituting an important revolution in theory. This is crucial to Coates's story about the influence of Wittgenstein on Keynes. But this attribution is totally at odds with Keynes's comments on these chapters, comments that not only reveal his attitudes towards his definitions but also seem a fair commentary on them.

> I have felt that these chapters were a great drag on getting on to the real business, and would perplex the reader quite unnecessarily with a lot of points which really do not matter to my proper theme (Keynes to Roy Harrod, 9 August 1935, quoted in (Keynes, 1971-1989, XIII: 537)).
>
> But the main point I would urge is that all this is *not* fundamental. *Being clear* is fundamental, but the choice of definitions of income and investment is not (Keynes to Dennis Robertson, 29 January 1935, quoted in (Keynes, 1971-1989, XIII, 495, italics in original)).

5 Keynes on Rules and Private Language

Had Keynes followed Wittgenstein in the ways suggested by either Bateman or Coates he would have been led into error. Fortunately he was not tempted. There was, however, one point on which Keynes clearly did not follow Wittgenstein, and sadly so for Wittgenstein was right. If Kripke (1982) is correct and this is the crucial point in the later Wittgenstein's thinking, Keynes's failure to observe it provides strong evidence that Wittgenstein's influence on him was at best slight.

Keynes, as we saw above, thought we dealt with uncertainty by assuming that the future would resemble the present. Call this Keynes's maxim. But this, points out Wittgenstein, gets us nowhere. We know that the future will resemble the present; what we don't know is how it will do so. Wittgenstein illustrates this with examples from mathematics and semantics, but we can apply it more broadly.

Say that a particle in a one-dimensional Euclidean space is now at position d, travelling at velocity v under acceleration a. Assuming things stay the same, where will the particle be in 1 unit of time? This question simply can't be answered, until we know what in what respect things will 'stay the same'. If it is in respect of position, the answer is d, in respect of velocity it is $d + v$, in respect of acceleration $d + v + a/2$. Perhaps our Newtonian intuitions make us prefer the second answer, perhaps not.

The same story applies in economics. When we assume things will stay the same, does that mean we are assuming the unemployment rate or the rate of change of the unemployment rate to be the same; real growth or nominal growth to be constant? At the level of the firm, we can ask whether Keynes's maxim would have us assume real or nominal profits to be constant, or perhaps the growth rate of real or nominal profits, or perhaps sales figures (real or nominal, absolute or variation), or perhaps one of the variables which play a role like acceleration (rate of change of sales growth)? In some computing firms we might even take some of the logarithmic variables (growth of logarithm of sales) to be the constant. We can't in consistency assume more than one or two of these variables to be unchanged, yet Keynes provides us with nothing to tell between them.

More importantly, it looks like Keynes hasn't even seen the problem. The mechanical example above looks very similar to some of the paradoxes of indifference (*TP*: Ch. 4). For example, in von Kries's cube factory example, we know that a factory makes cubes with side length between 0 and 2cm. If that's all we know, what should we say is the probability that the next cube's side length will be greater than 1cm? According to Laplace's principle of indifference we should divide the probabilities equally between the possibilities, which seems to give an answer of 1/2. However we could have set out the problem by saying that the volume of cubes produced is between 0 and $8cm^3$ and we want to know the probability the volume of the next cube is greater than $1cm^3$. Now the answer (to the same problem) looks to be 7/8. And if we set out the problem in terms of surface area we seem to get the answer 3/4. The conclusion is that the principle of indifference could only be saved if we have a small designated set of predicates to which we can exclusively apply it. But now it seems Keynes's maxim can only work if we have a small designated set of predicates to which we can exclusively apply it, and if we do that we can avoid the paradoxes of indifference. Keynes explicitly adopts his maxim to avoid the paradoxes of indifference (*GT*: 152). He would hardly have done this if he knew structurally similar problems beset the maxim as best the principle of indifference. As further evidence he just missed this point, note that while he was not averse to wielding philosophical tools in economic writing (like the paradoxes of indifference), Wittgenstein's point is not mentioned; not in the *GT*, not in any of its drafts and not in any of the correspondence after it was published.

For Kripke, this point is central to Wittgenstein's private language argument. All that we can know about the meaning of a word is how our community has used it in the past. We must assume they'll use it the same way in the future. But what is to count as using it the same way? *A priori* it looks like any usage of a word could count; the only thing that could make usage of a word wrong is the user has a different way of using the word 'the same way' to everyone else. Hence if there is no community to set such standards there are no bars on how words can be used. And if there are no such bars, there is nothing that can properly be called a language. Hence there can't be a private language.

Given the importance of that conclusion to Wittgenstein's later philosophy, if Kripke is even close to right in his reconstruction then it is central to the later Wittgenstein that Keynes's maxim is contentless. As Keynes clearly didn't think this (witness the central role it plays in summaries of the *GT* like Keynes 1937) he hasn't adopted a central tenet of the later Wittgenstein's work. This puts a rather heavy burden on those who would say he became a Wittgensteinian. The arguments presented so far do nothing to lift that burden.

Doing Philosophy With Words

Scott Soames has written two wonderfully useful books that will be valuable introductions to twentieth century philosophy. The books arose out of his well-received classes on the history of twentieth century history at Princeton, and will be valuable to anyone teaching similar courses. I shall be relying on them as I teach such a course at Cornell.

The books consist of detailed case studies of important twentieth-century works. They are best read alongside those original texts. Anyone who works through the canon in this way will have an excellent introduction to what twentieth century philosophers were trying to do. The selections are judicious, and while some are obvious classics some are rather clever choices of papers that are representative of the type of work being done at the time. And Soames doesn't just point to the most important works to study, but the most important sections of those works.

Soames's discussion of these pieces is always built around an analysis of their strengths and weaknesses. He praises the praiseworthy, but the focus, at least in the sections I'm discussing (ordinary language philosophy from Wittgenstein to Grice), is on where these philosophers go wrong. This is particularly so when the mistakes are representative of a theme. There are three main mistakes Soames finds in philosophers of this period. First, they rely logical positivism long after it had been shown to be unviable. Second, they disregard the principle that semantics should be systematic. Third, they ignore the distinction between necessity and a priority. All three constitute major themes of Soames's book, and indeed of twentieth century philosophy as Soames sees it.

These books concentrate, almost to a fault, on discussion of philosophers' published works, as opposed to the context in which they are written. Apart from occasionally noting that some books were released posthumously, we aren't told whether the philosophers who wrote them are alive, and only in one case are we told when a philosopher was born. This kind of external information does not seem important to Soames. He is the kind of historian who would prefer a fourth reading of Austin's published works to a first reading of his wartime diaries. And he'd prefer to spend the evening working on refutations, or charitable reformulations, of Austin's arguments to either. I'm mostly sympathetic to this approach; this is history of *philosophy* after all. We can leave discussions of the sociology of 1950s Oxford to those better qualified. But this choice about what to write about has consequences.

Most of Soames's chapters focus almost exclusively on a particular book or paper. The exceptions are like the chapter on *Sense and Sensibilia*, where Soames contrasts Austin's discussion with Ayer's response. We learn a lot about the most important works that way, but less about their intellectual environment. So the book doesn't have much by way of broad discussion about overall trends or movements. There's very little, for example, about who were the influencers and who the influenced. There's nothing about how anyone not called 'Wittgenstein' changed their positions in response to criticism. One assumes from the chronology that Ryle's influence on Austin was greater than Austin's influence on Ryle, for example, but Soames is silent on whether this is true.

† Penultimate draft only. Please cite published version if possible. Final version published in *Philosophical Studies* 135 (2007): 429-37. Thanks to David Chalmers, Michael Fara, John Fischer, Tamar Szabó Gendler, James Klagge, Michael Kremer, Ishani Maitra, Aidan McGlynn, Alva Noë, Jonathan Weinberg and Larry Wright.

Soames says at one point that, "[Ryle] was, along with Wittgenstein, J. L. Austin, and Paul Grice, one of the prime movers in postwar philosophy in England." (68). But we aren't really told why this is so, apart from the discussion of some prominent works of these four philosophers. (Perhaps Soames has taken the maxim *Show it, don't say it* rather completely to heart.) Nor are why told why the list includes those four, and not, say, Strawson or Geach or Anscombe. Actually Anscombe's absence reminds us that there is almost no discussion of women in philosophy in the book. That's not Soames fault, it's a reflection of a long-running systematic problem in philosophy that the discipline has a hard time recruiting and retaining women. Could some of that be traced back to what was going on in the ordinary language period? That kind of questions *can't* be addressed by the kind of history book that Soames has written, where the focus is on the best philosophical writing, and not on the broader philosophical community.

One of the other consequences of the format is that, by necessity, many important figures are left out, on pain of writing a fifteen-volume book. In the period under discussion here there was historically important work by (among many others) Nelson Goodman, Wilfrid Sellars and Roderick Chisholm, some of which connects up closely to the themes and interests of the ordinary language philosophers, but none of which is as much as mentioned. (Goodman is mentioned in the epilogue as someone Soames regrets not covering.)

Now this can't be a complaint about the book Soames has written, because it would have been impossible to cover any more figures than he did in the style and depth that he did. And it would have been impossible to tell in detail the story of how Ryle's impact on the philosophical world differed from Austin's, or of the painfully slow integration of women into the top echelons of philosophy, without making the book be even more monumental than it is. All we're left with is a half-hearted expression of regret that he didn't write a different *kind* of book, one that told us more about the forest, even as we value what he says about the tallest of the trees.

1 Grice and The End of Ordinary Language

There is one place where Soames stops to survey the field, namely his discussion of the impact of Grice's work on the ordinary language tradition. Soames argues that with Grice's William James lectures, the idea of ordinary language philosophy had "run their course". The position seems to be that Grice overthrew a paradigm that had been vibrant for two decades, but was running out of steam by the time of Grice's James lectures. How plausible is this?

The first step is to work out just what it was that Grice refuted. When summarising the ordinary language paradigm that he takes Grice to have overthrown, Soames is uncharacteristically harsh. In Soames's summary one of the characteristic activities of an ordinary language philosopher is "opportunistically assembling reminders about how philosophically significant words are used in ordinary settings" (216). That *may* be a fair enough description of *some* mid-century work, but it isn't a fair summary of the best of the work that Soames has spent the previous two hundred odd pages discussing. It all suggests that Grice didn't so much overthrow ordinary language philosophy as much as badly done ordinary language philosophy, and this category might not include Strawson, Ryle, Austin and so on.

More importantly, it isn't entirely clear just what it was Grice did that caused this paradigm shift. In Soames's telling it seems the development of the speaker meaning/semantic meaning distinction was crucial, but Austin at least already recognised this distinction, indeed appealed

to it twice in *Sense and Sensibilia*. Soames mentions the discussion on pages 89 to 91 of *Sense and Sensibilia* of phrases like "I see two pieces of paper", and there is also the intriguing discussion on pages 128-9 of the relation between *accurate* and *true* where Austin goes close to stating Grice's submaxim of concision.

The other suggestion is that Grice restored the legitimacy and centrality of systematic semantic theorising. It's true Grice did that, but this doesn't show we have to give up ordinary language philosophy unless it was impossible to be an ordinary language philosopher and a systematic semanticist. And it isn't clear that this really is impossible. It hardly seems *inconsistent* with the kind of philosophy Austin did (especially in his theory of perception) that one endorse a systematic semantic theory. (Though Austin *himself* rarely put forward systematic analyses.) Notably, there are plenty of very systematic formal semanticists who take Strawson's work on descriptions seriously, and try and integrate it into formal models. So we might wonder why Grice's work shouldn't have led to a kind of ordinary language philosophy where we paid more careful attention to system-building.

More broadly, we might wonder whether the ordinary language period really did end. The analysis of knowledge industry (strangely undiscussed in a work on *analysis* in the twentieth century) seemed to putter along much the same before and after the official demise of ordinary language philosophy. And there are affinities between the ordinary language philosophers and important contemporary research programs, e.g. the 'Canberra Plan' as described by Frank Jackson (1998). So perhaps before we asked who killed ordinary language philosophy (It was Professor Grice! In Emerson Hall!! With the semantics/pragmatics distinction!!!) we should have made sure there was a corpse. More on this point presently.

2 A Whig History?

One of the major themes of Soames's discussion is that there are some systematic problems in twentieth century philosophy that are righted by the heroes at the end of the story. I already mentioned the heroic role assigned to Grice. But the real star of the show is Kripke, who comes in as a deus ex machina at the end showing how different necessity and a priority are, and thereby righting all manner of grievous wrongs. That Kripke is an important figure in twentieth century philosophy is hardly a matter of dispute, but Soames does stretch a little to find errors for our hero to correct.

Some of the complaints about philosophers collapsing the necessary/a priori distinction do hit the target, but don't leave deep wounds in their victims. For instance, Soames quotes Ryle arguing (in *Dilemmas*) that perception cannot be a physiological process because if it were we couldn't *know* whether we saw a tree until we found out the result of complicated brain scans. Soames points out, perfectly correctly, that the seeing might be necessarily identical to the brain process even if we don't know, and even can't know without complicated measurements, whether they are identical. Soames is right that Ryle has made an epistemological argument here when a metaphysical argument was needed. But rewriting Ryle so he makes that metaphysical argument isn't hard. If my seeing the tree is necessarily identical to the brain process, and the brain process is (as Ryle and Soames seem to agree it is) individuated by the brain components that implement it, then I couldn't have seen the tree had one of the salient neurons in my brain been silently replaced with a functionally equivalent silicon chip. Since it *is* possible that I could have seen a tree even if a salient neuron was replaced with a functionally

equivalent silicon chip, the seeing and the brain process are not necessarily identical. So while Ryle might have slipped here, and Kripke's work does help us correct the slip, the consequences of this are basically verbal.

A more important charge of ignoring the necessary/a priori distinction comes in Soames's discussion of Wittgenstein's deflationism about philosophy. Here is the salient passage.

> His deflationary conception of philosophy is also consistent with, and even derivative from, his new ideas about meaning plus a set of unquestioned philosophical presuppositions he brings to the enterprise. The philosophical presuppositions include the then current and widespread assumptions that (i) that philosophical theses are not empirical, and hence must be necessary and a priori, and (ii) that the necessary, the a priori and the analytic are one and the same. Because he takes these assumptions for granted, he takes it for granted that if there are any philosophical truths, they must be analytic (29).

This seems to me to be mistaken twice over.

First, it isn't clear to me that there is *any* appeal to concepts of necessity in the passages in Wittgenstein Soames is summarising here, and metaphysical necessity simply doesn't seem to have been a major interest of Wittgenstein's. Wittgenstein does appear to reason that if a proposition is not empirical it is a priori, but that inference doesn't go via claims about necessity, and isn't shown to be fallacious by any of Kripke's examples.

Second, it simply isn't true that philosophers in Wittgenstein's time took for granted that the analytic and the a priori were one and the same. To be sure, many philosophers in the early twentieth century (including many argue the younger Wittgenstein) argued against Kant's claim that they are distinct, but this isn't quite the same as taking for granted they are identical. And there are a few places where Wittgenstein appears to accept that some propositions are synthetic a priori. For example in *Remarks on the Foundations of Mathematics* he says it is synthetic a priori that there is no reddish green, (Part III, para 39) and goes on to say this about primes.

> The distribution of primes would be an ideal example of what could be called synthetic a priori, for one can say that it is at any rate not discoverable by an analysis of the concept of a prime number. (Wittgenstein, 1956, Part III, para 42)

Now it is far from obvious what the connection is between remarks such as these and the remarks about the impossibility of philosophical theses in the *Investigations*. Indeed it is not obvious whether Wittgenstein really believed in the synthetic a priori at any stage of his career. But given his lack of interest in metaphysical necessity, and openness to the possibility of synthetic a priori claims, it seems unlikely that he was, tacitly or otherwise, using the argument Soames gives him to get the deflationary conclusions.[1]

[1] I'm grateful to many correspondants for discussions about Wittgenstein. They convinced me, inter alia, that it would be foolish of me to commit to strong views of any kind about the role of the synthetic a priori in Wittgenstein's later thought, and that the evidence is particularly messy because Wittgenstein wasn't as centrally concerned with these concepts as we are.

3 Getting the Question Right

As I mentioned above, Soames's is the kind of history that focuses on the works of prominent philosophers, rather than their historical context. There's much to be gained from this approach, in particular about what the greats can tell us about pressing philosophical questions. But one of the costs is that in focussing on what they say about *our* questions, we might overlook *their* questions. In most cases this is a trap Soames avoids, but in the cases of Austin and Ryle the trap may have been sprung.

Soames sees Austin in *Sense and Sensibilia* as trying to offer us a new argument against radical scepticism.

> Austin's ultimate goal is to undermine the coherence of skepticism. His aim is not just to show that skepticism is unjustified, or implausible, or that it is a position no one has reason to accept. Rather, his goal is to prevent skepticism from getting off the ground by denying skeptics their starting point. (173-4)

But we don't get much of an interpretative argument that this is really Austin's goal. Indeed, Soames concedes that Austin "doesn't always approach these questions directly" (172). I'd say he does very little to approach them at all. To be sure, many contemporary defenders of direct realism are interested in its anti-sceptical powers, but there's little to show *Austin* was so moved. Scepticism is not a topic that even arises in *Sense and Sensibilia* until the chapter on Warnock, after Austin has finished with the criticism of Ayer that takes up a large part of the book. And Soames doesn't address the question of how to square the somewhat dismissive tone Austin takes towards scepticism in "Other Minds" with the view here propounded that Austin put forward a fairly radical theory of perception as a way of providing a new answer to the sceptic.

If Austin wasn't trying to refute the sceptic, what was he trying to do? The simplest explanation is that he thought direct realism was true, sense-data theories were false, and that "there is noting so plain boring a the constant repetition of assertions that are not true, and sometimes no even faintly sensible; if we can reduce this a bit, it will all be to the good." (Austin, 1962, 5) I'm inclined to think that in this case the simplest explanation is the best, that Austin wrote a series of lectures on perception because he was interested in the philosophy of perception. Warnock says that "Austin was genuinely shocked by what appeared to his eye to be recklessness, hurry, unrealism, and inadequate attention to truth" (Warnock, 1989, 154) and suggests this explained not only why Austin wrote the lectures but their harsh edge.

There is one larger point one might have wanted to make out of a discussion of direct realism, or that one might have learned from a discussion of direct realism, that seems relevant to what comes later in Soames's book. If we really see objects, not sense-data, then objects are constituents of intentional states. That suggests that public objects might be constituents of other states, such as beliefs, and hence constituents of assertions. Soames doesn't give us a discussion of these possible historical links between direct realism and direct reference, and that's too bad because there could be some fertile ground to work over here. (I'm no expert on the history of the 1960s, so I'm simply guessing as to whether there is a historical link between direct realism and direct reference to go along with the strong philosophical link between the two. But it would be nice if Soames has provided an indication as to whether those guesses were likely to be productive or futile.)

Soames gives us no inkling of where theories of direct reference came from, save from the brilliant mind of Kripke. Apart from the absence of discussion of any connection between direct realism and direct reference, there's no discussion of the possible connections between Wittgenstein's later theories and direct reference, as Howard Wettstein (2004) has claimed exist. And there's no discussion of the (possibly related) fact that Kripke was developing the work that went into *Naming and Necessity* at the same time as he was lecturing and writing on Wittgenstein, producing the material that eventually became *Wittgenstein on Rules and Private Language*. Kripke is presented here as the first of the moderns[2], and in many ways he is, but the ways in which he is the last (or the latest) of the ordinary language philosophers could be a very valuable part of a history of philosophy.[3]

Matters are somewhat more difficult when it comes to Ryle's *The Concept of Mind*. Ryle predicted that he would "be stigmatised as 'behaviourist"' (Ryle, 1949, 327) and Soames obliges, and calls him a verificationist to boot.

> If beliefs and desires were private mental states [says Ryle], then we could never observe the beliefs and desires of others. But if we couldn't observe them, then we couldn't know that they exist, [which we can.] … This argument is no stronger than verificationism in general, which by 1949 when *The Concept of Mind* was published, had been abandoned by its main proponents, the logical positivists, for the simple reason that every precise formulation of it had been decisively refuted (97-8).

But Ryle's position here isn't verificationism at all, it's abductophobia, or fear of inference to underlying causes. Ryle doesn't think the claim of ghosts in the machine is *meaningless*, he thinks it is false. The kind of inference to underlying causes he disparages here is *exactly* the kind of inference to unobservables that paradigm verificationists, especially Ayer, go out of their way to *allow*, and in doing so buy all end of trouble.[4] And abductophobia is prevalent among many contemporary *anti*-verificationists, particularly direct realists such as McDowell (1996), Brewer (1999) and Smith (2003) who think that if we don't directly observe beer mugs we can never be sure that beer mugs exist. I basically agree with Soames that Ryle's argument here (and the same style of argument recurs repeatedly in *The Concept of Mind*) is very weak, but it's wrong to call it verificationist.

The issue of behaviourism is trickier. At one level Ryle surely is a behaviourist, because whatever *behaviourism* means in philosophy, it includes what Ryle says in *The Concept of Mind*. Ryle is the reference-fixer for at least one disambiguation of *behaviourist*. However we label Ryle's views though, it's hard to square what he says his aims are with the aims Soames attributes to him. In particular, consider Soames's criticism of Ryle's attempt to show that we don't need

[2]The first of what David Armstrong (2000) has aptly called "The Age of Conferences".

[3]Just in case this gets misinterpreted, what I'm suggesting here is that Kripke (and his audiences) might have been influenced in interesting ways by philosophy of the 1950s and 1960s, *not* that Kripke took his ideas from those philosophers. The latter claim has been occasionally made, but on that 'debate' (Soames, 1998b,a) I'm 100% on Soames's side.

[4]It would be particularly poor form of me to use a paradigm case argument without discussing Soames's very good dissection of Malcolm's paradigm case argument in chapter 7 of his book. So let me note my gratitude as a Cornellian for all the interesting lines of inquiry Soames finds suggested in Malcolm's paper – his is a paradigm of charitable interpretation, a masterful discovery of wheat where I'd only ever seen chaff.

to posit a ghost in the machine to account for talk of intelligence. (Soames is discussing a long quote from page 47 of *The Concept of Mind*.)

> The description Ryle gives here is judicious, and more or less accurate. But it is filled with words and phrases that seem to refer to causally efficacious internal mental states—*inferring*, *thinking*, *interpreting*, *responding to objections*, *being on the lookout for this*, *making sure not to rely on that*, and so on. Unless all of these can be shown to be nothing more than behavioral dispositions, Ryle will not have succeeded in establishing that to argue intelligently is simply to manifest a variety of purely behavioral dispositions. (106)

And Soames immediately asks

> So what are the prospects of reducing all this talk simply to talk about what behavior would take place in various conditions? (106)

The answer, unsurprisingly, is that the prospects aren't good. But why this should bother *Ryle* is never made clear. For Ryle only says that when we talk of mental properties we talk about people's dispositions, not that we talk about their *purely behavioural* dispositions. The latter is Soames's addition. It is rejected more or less explicitly by Ryle in his discussion of knowing how. "Knowing *how*, then, is a disposition, but not a single-track disposition like a reflex or a habit … its exercises can be overt or covert, deeds performed or deeds imagined, words spoken aloud or words heard in one's head, pictures painted on canvas or pictures in the mind's eye." (1949, 46-47). Nor should Ryle feel compelled to say that these dispositions are behavioural, given his other theoretical commitments.

Ryle is opposed in general to talk of 'reduction' as the discussion of mechanism on pages 76ff shows. To be sure there he is talking about reduction of laws, but he repeatedly makes clear that he regards laws and dispositions as tightly connected (1949, 43, 123ff) and suggests that we use mental concepts to signal that psychological rather than physical laws are applicable to the scenario we're discussing (167). Moreover, he repeatedly talks about mental events for which it is unclear there is any kind of correlated *behavioural* disposition, e.g. the discussion of Johnson's stream of consciousness on page 58 and the extended discussion of imagination in chapter 8. Ryle's claim that "Silent soliloquy is a form of pregnant non-sayings" (269) hardly looks like the claim of someone who wanted to reduce all mental talk to behavioural dispositions, unless one leans rather hard on 'pregnant'. But we aren't told whether Soames leans hard on this word, for he never quite tells us why he thinks all the dispositions that Ryle considers must be behavioural dispositions, rather than (for example) dispositions to produce other dispositions.

To be sure, from a modern perspective it is hard to see where the space is that Ryle aims to occupy. He wants to eliminate the ghosts, so what is left for mind to be but physical stuff, and what does physical stuff do but behave? He's not an eliminativist, so he's ontologically committed to minds, and he hasn't left anything for them to be but behavioural dispositions. So we might see it (not unfairly) but that's not how Ryle sees it.[5] Soames sees Ryle as an ancestor

[5]Of course he *couldn't* have seen it that way since in 1949 he wouldn't have had the concept of ontological commitment.

of a reductive materialist like David Lewis, and a not very successful one at that. But the Ryle of *The Concept of Mind* has as much in common with non-reductive materialists, especially when he says that "not all questions are physical questions" (1949, 77), insists that "men are not machines, not even ghost-ridden machines" (1949, 81) and describes Cartesians rather than mechanists as "the better soldiers" (1949, 330) in the war against ignorance. Perhaps a modern anti-dualist should aim for a reduction of the mental to the physical, but Ryle thought no such reduction was needed to give up the ghost, and the historian should record this.

4 Conclusion

As I said at the top, Soames has written two really valuable books. For anyone who wants to really understand the most important philosophical work written between 1900 and 1970, reading through the classics while constantly referring back to Soames's books to have the complexities of the philosophy explained will be immensely rewarding. Those who do that might feel that the people who skip reading the classics and just read Soames's books get an unreasonably large percentage of the benefits they've accrued. As noted once or twice above I have some quibbles with some points in Soames's story, but that shouldn't let us ignore what a great service Soames has provided by providing these surveys of great philosophical work.

In Defense of a Kripkean Dogma

Jonathan Ichikawa, Ishani Maitra, Brian Weatherson

In "Against Arguments from Reference" (Mallon et al., 2009), Ron Mallon, Edouard Machery, Shaun Nichols, and Stephen Stich (hereafter, MMNS) argue that recent experiments concerning reference undermine various philosophical arguments that presuppose the correctness of the causal-historical theory of reference. We will argue three things in reply. First, the experiments in question—concerning Kripke's Gödel/Schmidt example—don't really speak to the dispute between descriptivism and the causal-historical theory; though the two theories are empirically testable, we need to look at quite different data than MMNS do to decide between them. Second, the Gödel/Schmidt example plays a different, and much smaller, role in Kripke's argument for the causal-historical theory than MMNS assume. Finally, and relatedly, even if Kripke *is* wrong about the Gödel/Schmidt example—indeed, even if the causal-historical theory is not the correct theory of names for some human languages—that does not, contrary to MMNS's claim, undermine uses of the causal-historical theory in philosophical research projects.

1 Experiments and Reference

MMNS start with some by now famous experiments concerning reference and mistaken identity. The one they focus on, and which we'll focus on too, is a variant of Kripke's Gödel/Schmidt example. Here is the question they gave to subjects.

> Suppose that John has learned in college that Gödel is the man who proved an important mathematical theorem, called the incompleteness of arithmetic. John is quite good at mathematics and he can give an accurate statement of the incompleteness theorem, which he attributes to Gödel as the discoverer. But this is the only thing that he has heard about Gödel. Now suppose that Gödel was not the author of this theorem. A man called "Schmidt" whose body was found in Vienna under mysterious circumstances many years ago, actually did the work in question. His friend Gödel somehow got hold of the manuscript and claimed credit for the work, which was thereafter attributed to Gödel. Thus he has been known as the man who proved the incompleteness of arithmetic. Most people who have heard the name 'Gödel' are like John; the claim that Gödel discovered the incompleteness theorem is the only thing they have ever heard about Gödel. When John uses the name 'Gödel,' is he talking about:
>
> (A) the person who really discovered the incompleteness of arithmetic? or
> (B) the person who got hold of the manuscript and claimed credit for the work? (MMNS 2009: 341)

† Penultimate draft only. Please cite published version if possible. Final version forthcoming in *Philosophy and Phenomenological Research*.

The striking result is that while a majority of American subjects answer (B), consistently with Kripke's causal-historical theory of names, the majority of Chinese subjects answer (A).[1] To the extent that Kripke's theory is motivated by the universality of intuitions in favour of his theory in cases like this one, Kripke's theory is undermined.

There are now a number of challenges to this argument in the literature. Before developing our own challenge, we'll briefly note five extant ones, which all strike us as at least approximately correct.

(1) Kripke's theory is a theory of *semantic* reference. When asked who John is talking about, it is natural that many subjects will take this to be a question about *speaker* reference. And nothing in Kripke's theory denies that *John* might refer to the person who proved the incompleteness of arithmetic, even if his word refers to someone else. (Ludwig, 2007; Deutsch, 2009)

(2) Kripke's argument relies on the *fact* that 'Gödel' refers to Gödel, not to the universality or otherwise of intuitions about what it refers to. That some experimental subjects don't appreciate this fact doesn't make it any less of a fact. (Deutsch, 2009)

(3) If the subjects genuinely were descriptivists, it isn't clear how they could make sense of the vignette, since the name 'Gödel' is frequently used in the vignette itself to refer to the causal origin of that name, not to the prover of the incompleteness or arithmetic.[2] On a related point, Martí doesn't mention this, but subjects who aren't descriptivists should also object to the vignette, since in the story John doesn't learn Gödel proved the incompleteness of arithmetic, at least not if 'learn' is factive. (Martí, 2009)

(4) The experiment asks subjects for their judgments about a metalinguistic, and hence somewhat theoretical, question about the mechanics of reference. It's better practice to observe how people actually refer, rather than asking them what they think about reference. (Martí, 2009; Devitt, 2010)

(5) Intuitions about the Gödel/Schmidt case play at best a limited role in Kripke's broader arguments, so experimental data undermining their regularity do not cast serious doubt on Kripke's theory of reference. (Devitt, 2010)

We think challenges (1)-(3) work. Something like (4) should work too, although it requires some qualification. Consider, for instance, what happens in syntax. It's true, of course, that we don't go around asking ordinary speakers whether they think *Lectures on Government and*

[1] Note that a causal descriptivist about names will also say that the correct answer to this question is (B). So the experiment isn't really testing descriptivism as such versus Kripke's causal-historical theory, but some particular versions of descriptivism against Kripke's theory. These versions of descriptivism say that names refer to the satisfiers of (generally non-linguistic) descriptions that the name's user associates with the name. One such version is 'famous deeds' descriptivism, and the descriptions MMNS use are typically famous deeds; nevertheless, that seems inessential to their experiments. When we use 'descriptivism' in this paper, we'll mean any such version of descriptivism. Thanks here to an anonymous referee.

[2] This objection relies on an empirical assumption that may be questionable. It assumes that the subject of the experiment associates the same description with 'Gödel' as John does. A subject who (a) is a descriptivist and (b) associates with the name 'Gödel' the description 'the man who proved the compatibility of time travel and general relativity', can also make sense of the vignette, *contra* Martí. So perhaps the objection could be resisted. But we think this empirical assumption is actually fairly plausible. Unless the experimental subjects were being picked from a very biased sample, the number of subjects who are familiar with Gödel's work on closed time-like curves is presumably vanishingly small! We're grateful here to an anonymous referee.

Binding was an advance over *Aspects*. Or, if we did, we wouldn't think it had much evidential value. But that's not because ordinary speaker judgments are irrelevant to syntax. On the contrary, judgments about whether particular strings constitute well-formed sentences are an important part of our evidence.[3] But they are not our only evidence, or even our primary evidence; we also use corpus data about which words and phrases are actually used, and many syntacticians take such usage evidence to trump evidence from metasemantic intuitions.[4] Even when we do seek such intuitive answers, perhaps because there isn't enough corpus data to settle the usage issue, the questions might be about cases that are quite different to the cases we primarily care about. So we might ask a lot about speakers' judgments concerning questions even if we care primarily about the syntax of declarative sentences.

If what Kripke says in *Naming and Necessity* (hereafter, NN) is right, then we should expect something similar in the case of reference. Kripke anticipates that some people will disagree with him about some of the examples, and offers a few replies. (Our discussion here largely draws on footnote 36 of NN.) Part of his reply is a version of point 1 above; those disagreements may well be over speaker reference, not semantic reference. That reply is correct; it's hard for us to hear a question about who someone is talking about as anything but a question about speaker reference. He goes on to note that his theory makes empirical predictions about how names are used.

> If I mistake Jones for Smith, I may *refer* (in an appropriate sense) to Jones when I say that Smith is raking the leaves … Similarly, if I erroneously think that Aristotle wrote such-and-such passage, I may perhaps sometimes use 'Aristotle' to *refer* to the actual author of the passage … In both cases, I will withdraw my original statement, and my original use of the name, if apprised of the facts. (NN 86n)

This seems entirely right. There's some sense in which John, in MMNS's vignette, is referring to Gödel and some sense in which he's referring to Schmidt. Just thinking about the particular utterance he makes using 'Gödel' won't help much in teasing apart speaker reference and semantic reference. What we should look to are patterns of—or if they're not available, intuitions about—withdrawals of statements containing disputed names. To use the example Kripke gives here, consider a speaker who (a) associates with the name 'Aristotle' only the description 'the author of *The Republic*', (b) truly believes that a particular passage in *The Republic* contains a quantifier scope fallacy, and (c) is a descriptivist. She might say "Aristotle commits a quantifier

[3] This point suggests Martí's criticism of MMNS as stated overshoots. She wants to dismiss arguments from metalinguistic intuitions altogether. But intuitions about well-formedness *are* metalinguistic intuitions, and they are a key part of the syntactician's toolkit. Martí concedes something like this point, but claims that the cases are not on a par, because syntax concerns a normative issue and reference does not. We're quite suspicious that there's such a striking distinction between the kind of subject-matter studied by syntacticians and semanticists. Devitt's version of this point is more modest and does not obviously commit to this exaggeration.

[4] Here's one example where testing intuitions and examining the corpus may lead to different answers. Many people think, perhaps because they've picked up something from a bad style guide, that the sentence 'Whenever someone came into Bill's shop, he greeted them with a smile', contains one or two syntactic errors. (It uses a possessive as the antecedent of a pronoun, and it uses 'them' as a bound singular variable.) Even if most subjects in a survey said such a sentence was not a well-formed sentence of English, corpus data could be used to show that it is. Certainly the existence of a survey showing that users in, say, Scotland and New Jersey give different answers when asked about whether the sentence is grammatical would not show that there's a syntactic difference between the dialects spoken in Scotland and New Jersey. You'd also want to see how the sentences are *used*.

scope fallacy in this passage." When she's informed that the passage was written by Plato, she'll no longer utter those very words, but she'll still insist that the sentence she uttered was literally true. That's because she'll claim that in that sentence 'Aristotle' just referred to the author of the passage, and that person did commit a quantifier scope fallacy. A non-descriptivist will take back the claim expressed, though she might insist that what she *intended* to say was true.

So to show that subjects in different parts of the world really have descriptivist intuitions about the Gödel/Schmidt case, we might ask about whether they think John should withdraw, or clarify, his earlier statements if apprised of the facts. Or we might ask whether they would withdraw, or clarify, similar statements they had made if apprised of the facts. Or, even better, we might test whether in practice people in different parts of the world really do withdraw their prior claims at different rates when apprised of the facts about a Gödel/Schmidt case. Kripke is right that given descriptivism, a speaker shouldn't feel obliged to withdraw the original statement when apprised of the facts, but given the causal-historical theory, they should. So there are experiments that we could run which would discriminate between descriptivist and causal-historical approaches, but we don't think the actual experiment MMNS run does so.

In its broad terms, we agree with Devitt's challenge (5), although we understand the role of the Gödel/Schmidt case rather differently than he does. We turn now to this question.

2 Gödel's Role in Naming and Necessity

In the first section we argued that the experimental data MMNS offer do not show that the correct account of the Gödel/Schmidt example is different in different dialects. In this section we want to argue that there's very little one *could* show about the Gödel/Schmidt example that would bear on the broader question of what the correct theory of reference is. To see this, let's review where the Gödel/Schmidt example comes up in *Naming and Necessity*.

In the first lecture, Kripke argues, via the modal argument, that names can't be synonymous with descriptions. The reason is that in modal contexts, substituting a name for an individuating description alters truth values. So a pure descriptivism that treats names and descriptions as synonymous is off the table. What's left, thinks Kripke, is what Soames calls "weak descriptivism" (Soames, 2003, Volume II, 356). This is the view that although names are not synonymous with descriptions, and do not abbreviate descriptions, they do have their reference fixed by descriptions. Here is the way Kripke introduces the picture that he is attacking.

> The picture is this. I want to name an object. I think of some way of describing it uniquely and then I go through, so to speak, a sort of mental ceremony: By 'Cicero' I shall mean the man who denounced Cataline … [M]y intentions are given by first, giving some condition which uniquely determines an object, then using a certain word as a name for the object determined by this condition. (NN 79)

The Gödel/Schmidt example, or at least the version of it that MMNS discuss, comes up in Kripke's attack on one of the consequences of this picture of naming. (A variant on the example, where no one proves the incompleteness of arithmetic, is used to attack another consequence of the theory.) So the role of the Gödel/Schmidt example is to undermine this picture of names and naming.

But note that it is far from the only attack on this picture. Indeed, it is not even the first attack. Kripke's first argument is that for most names, most users of the name cannot give an individuating description of the bearer of the name. In fact, those users cannot even give a description of the bearer that is individuating *by their own lights*. The best they can do for 'Cicero' is 'a Roman orator' and the best they can do for 'Feynman' is 'a famous physicist'. (NN 81) But it isn't that these users think that there was only one Roman orator, or that there is only one famous physicist. It's just that they don't know any more about the bearers of these names they possess. The important point here is that Kripke starts with some examples where the best description a speaker can associate with a name is a description that isn't individuating *even by the speakers' own lights*. And he thinks that descriptivists can't explain how names work in these cases.

Now perhaps we'll get new experimental evidence that even in these cases, some experimental subjects have descriptivist intuitions. Some people might intuit that if a speaker does not know of any property that distinguishes Feynman from Gell-Mann, their name 'Feynman' is indeterminate in reference between Feynman from Gell-Mann. We're not sure what such an experiment would tell us about the metaphysics of reference, but maybe someone could try undermining Kripke's argument this way. But that's not what MMNS found; their experiments don't bear on what Kripke says about 'Feynman', and hence don't bear on his primary argument against weak descriptivism.

Some philosophers will hold that although the picture Kripke describes here, i.e., weak descriptivism, can't be right in general for Feynman/Gell-Mann reasons, it could be true in some special cases. We agree. So does Kripke. The very next sentence after the passage quoted above says, "Now there may be some cases in which we actually do this." (NN 79) And he proceeds to describe three real life cases (concerning 'Hesperus', 'Jack the Ripper' and 'Neptune') where the picture is plausibly correct. But he thinks these cases are rare. In particular, we shouldn't think that the existence of an individuating description is sufficient reason to believe that we are in such a case. That, at last, is the point of the Gödel/Schmidt example. His conclusion from that example is that weak descriptivism isn't correct even in those special cases of names where the speaker possesses a description that she *takes* to be individuating.[5]

Michael Devitt (2010) also argues that MMNS exaggerate the importance of the Gödel/Schmidt case. He identifies a number of Kripke's other arguments (including the Feynman one we mention) that he takes to be more central, and, like us, he argues that MMNS's results do not cast doubt on these arguments. We agree, noting only two points of difference. First, as suggested above, although the Gödel/Schmidt case is not the only or the most central motivation for Kripke's theory of reference, we do think that it plays a distinctive role, compared with that of, for instance, the Feynman case. It refutes even the weak version of weak descriptivism according to which, in the special case in which subjects do possess individuating descriptions, those descriptions determine reference. We think the Gödel/Schmidt case (together

[5]The Gödel/Schmidt example is also distinctive in another way, in that the description in question actually applies to the referent of the name, and indeed speakers actually know this. But the flow of the text around the example (especially on page 84) suggests Kripke intends the example to make the same point as is made by other examples, such as the Peano/Dedekind case (in which the possessed description doesn't actually apply to the referent of the name). So this is probably not crucial to the point the example makes. We'll return below to the issue of just what this example shows. The key point is that the more distinctive the example is, the *less* that would follow if Kripke were wrong about the example; he might only be wrong about examples with just those distinctive features.

with the Peano/Dedekind case) form the basis of the only argument in *Naming and Necessity* against this weak weak descriptivism. (On a closely related point, we, unlike Devitt, take the Gödel/Schmidt case to be addressing a quantitative question about how common descriptive names are, not the qualitative question about whether the causal-historical theory is true at all; we'll expand on this point below.) Second, Devitt expresses some scepticism about the Gödel/Schmidt judgment on the grounds that the relevant case is somewhat 'fanciful'—actual cases, Devitt suggests, are better to be trusted. While there is surely some truth in the suggestion that intuitions about esoteric and complicated cases can be less trustworthy than those about everyday ones, we see little reason for concern in this instance; the Gödel case does not describe a scenario we should expect to find trouble thinking about.

Our reconstruction of the structure of Kripke's argument should make it clear how *unimportant* the Gödel/Schmidt example is to the broader theoretical questions. If Kripke were wrong about the Gödel/Schmidt case, that would at most show that there are a few more descriptive names than we thought there were. But since the existence of some descriptive names is consistent with the causal-historical theory of reference, the existence of a few more is too. All the Gödel/Schmidt example is used for in *Naming and Necessity* is to show that the number of descriptive names in English is not just small, it is *very* small. But the truth of the causal-historical theory of reference doesn't turn on whether there are few descriptive names, or very few descriptive names.

Once we see that the Gödel/Schmidt example concerns a quantitative question (are descriptive names rare or very rare?) rather than a qualitative question (is the causal-historical theory correct?), we can see some limitations of the experiment MMNS rely on. The case that MMNS describes to their subjects has several distinctive features, and it isn't clear that we'd be justified in drawing conclusions from it about cases that lack those features. Here is one such feature. The subject of the vignette (John) acquires the name 'Gödel' at the same time as he acquires an individuating description of Gödel. Suppose it turned out that, in some dialects at least, that would be sufficient for the name to be a descriptive name; i.e., for it to be a name whose reference is fixed by a description somehow attached to that name. If this conjecture is true, then descriptive names are a little more common than Kripke thinks they are, but not a lot more common. Now we don't actually think this conjecture is true. And for the reasons given in section 1 we don't think this experiment is evidence for it. What we do think is that (a) it's hard to see how studying reactions to cases like the Gödel/Schmidt example could show more than that some such claim about the prevalence of descriptive names is true, and (b) such claims are not inconsistent with the causal-historical theory.

We've argued that even if Kripke is wrong about the Gödel/Schmidt example, that doesn't undermine the arguments for the main conclusions of *Naming and Necessity*. A natural inference from this is that experiments about the Gödel/Schmidt example can't undermine those conclusions. We think the natural inference is correct. A referee has suggested that this is too quick. After all, if we have experimental evidence that Kripke is wrong about the Gödel/Schmidt case, we might have some grounds for suspicion about the other cases that Kripke uses in the arguments for more central conclusions. That is, if MMNS are right about the Gödel/Schmidt case, that doesn't give us a *deductive* argument against the other anti-descriptivist moves, but it might give us an *inductive* argument against them. This is an important worry, but we think it can be adequately responded to.

The first thing to note is that it would be foolish to fall back to a general scepticism about human judgment just because people disagree in their intuitive reactions to some tricky cases. This point is well argued by Timothy Williamson in his (2007, Ch. 6). If there's a worry here, it must be because the evidence about the Gödel/Schmidt example supports a more modest generalisation about judgments about cases, but that generalisation is nevertheless strong enough to undermine Kripke's other arguments. We doubt such a generalisation exists.

It can't be that the experiments about the Gödel/Schmidt example show that intuitive judgments about reference are systematically mistaken. Most of our intuitions in this field are surely correct. For instance, our intuitions that 'Kripke' refers to Kripke and not Obama, and that 'Obama' refers to Obama and not Kripke, are correct. (And experiments like the ones MMNS ran don't give us any reason at all to doubt that.) And we could produce many more examples like that. At most, the experiments can show us that there are spots of inaccuracy in a larger pool of correct judgments.

It might be argued that we should be sceptical of intuitions about reference in counterfactual cases. The correct judgments cited in the previous paragraph are all about real cases, but the Gödel/Schmidt example is not a real case. Now we don't think that the experiments do undermine all intuitions about reference in counterfactual cases, but even if they did, that wouldn't affect the Kripkean argument. That's because the central argument against descriptivism at the start of Lecture II involves real cases. The heavy lifting is done by cases where speakers don't think they have an individuating description to go along with names they use (e.g., 'Feynman' and 'Gell-Mann'), or they believe they have an individuating description, but that description involves some kind of circularity (e.g., 'Einstein', 'Cicero'). It seems to us that these cases are much more like the cases where we know people have accurate intuitions about reference (e.g., 'Obama' refers to Obama), than they are like cases where there is some dispute about their accuracy (e.g., 'Gödel' would refer to Gödel even if Schmidt had proved the incompleteness of arithmetic). So there's no reason to doubt the intuitions that underlie these central Kripkean arguments. And so there's no reason from these experiments to doubt the anti-descriptivist conclusions Kripke draws from them.

3 Reference in Philosophy

If the data about the Gödel/Schmidt example don't undermine the causal-historical theory of reference, then presumably they don't undermine philosophical uses of that theory. But we think MMNS overstate the role that theories of reference play in philosophical theorising, and we'll end by saying something about this.

One simple reaction to MMNS's argument is to say that at most they show that the causal-historical theory of reference is not true of some dialects. But, a philosopher might say, they are not writing in such a dialect, and the causal-historical theory is true of their dialect. And that's all they needed for their argument. MMNS anticipate this objection, and reply to it in section 3.3 of their paper. The reply is, in essence, that such a picture would make a mess of communication. If we posit dialectical variation to explain different reactions to the Gödel/Schmidt example, and to other examples, then we cannot know what dialect someone is speaking without knowing how they respond to these examples. And plainly we don't need to quiz people in detail about philosophical examples in order to communicate with them.

We offer three replies.

First, at least one of us is on record raising in principle suspicions about this kind of argument Maitra (2007). The take-home message from that paper is that communication is a lot easier than many theorists have supposed, and requires much less pre-communicative agreement. It seems to us that the reply MMNS offer here is susceptible to the arguments in that paper, but for reasons of space we won't rehearse those arguments in detail.

Second, it's one thing to think that variation in *reference* between dialects leads to communication breakdown, it's another thing altogether to think that variation in *meta-semantics* leads to such breakdown. A little fable helps make this clear. In some parts of Melbourne, 'Gödel' refers to Gödel because of the causal chains between the users of the name and the great mathematician. In other parts, 'Gödel' refers to Gödel because the speakers use it as a descriptive name, associated with the description 'the man who proved the incompleteness of arithmetic'. Kevin doesn't know which area he is in when he sees a plaque over a door saying "Gödel lived here". It seems to us that Kevin can understand the sign completely without knowing how 'Gödel' got its reference. Indeed, he even knows what proposition the sign expresses. So meta-semantic variation between dialects need not lead to communicative failure, even when hearers don't know which dialect is being used.

Third, if MMNS's argument succeeds, it seems to us that it shows descriptivist theories, including the weak weak descriptivism that Kripke is arguing against with the Gödel/Schmidt example, are doomed. (The arguments in this paragraph are not original. Similar arguments are used frequently in, e.g., Fodor and Lepore (1992).) It's a platitude that different people know different things. Barring a miracle, that means different people will associate different descriptions with different names. If there is widespread use of descriptive names, that means there will be widespread differences in which descriptions are associated with which names. And that will produce at least as much communicative difficulty as having some people be causal-historical theorists and some people be descriptivists. In short, if MMNS's argument against 'referential pluralism' is sound, there is an equally sound argument against descriptivism. And note that this argument doesn't rely on any thought experiments about particular cases. It doesn't even rely on thought experiments about names like 'Einstein', where there isn't any evidence that Kripke is wrong about how those names work.

Dialectically, the situation is this. MMNS have offered an argument from the possibility of communicating under conditions of ignorance about one's interlocutor's knowledge. Similar arguments have been offered against descriptivism. If such arguments are successful, then descriptivism is false, and there's no problem with philosophers making arguments from the falsity of descriptivism. If such arguments are unsuccessful, then MMNS haven't shown that it is wrong for philosophers to assume that the causal-historical theory is the right theory for *their* dialect, even if some other people are descriptivists. And, as MMNS concede, as long as the philosophers themselves speak a causal-historical theory dialect, the uses of the causal-historical theory in philosophy seem appropriate. The only way this argument could fail is if MMNS's argument from the possibility of communicating under conditions of ignorance about one's interlocutor's knowledge is stronger than the analogous arguments against descriptivism. But we see no reason to believe that is so. If anything, it seems like a weaker argument, because of the considerations arising from our fable about Kevin and the 'Gödel lived here' sign.

So we don't think MMNS have a good reply to the philosopher who insists that they only need the causal-historical theory to be true of *their* dialect. But in fact we think that philosophers rarely even assume that much.

Let's consider one of the examples that they cite: Richard Boyd's use of the causal-historical theory of reference in developing and defending his version of "Cornell Realism" in his (1988). Here's one way one could try and argue for moral realism from the causal-historical theory.

1. The causal-historical theory of reference is the correct theory of reference for all words in all dialects (or at least our dialect).
2. So, it is the correct theory for 'good'.

But that's not Boyd's actual argument. And that's a good thing, because the first premise is implausible. Someone defending it has to explain descriptive names like 'Neptune', logical terms like 'and', empty predicates like 'witch', and so on. And Boyd's not in that business. His argument is subtler. Boyd uses the causal-historical theory for two purposes. First, he uses the development of a naturalistically acceptable theory of reference as part of a long list of developments in post-positivist philosophy that collectively constitute a "distinctively realist conception of the central issues in the philosophy of science" (Boyd, 1988, 188). Second, he uses the causal-historical theory of reference, as it applies to natural kind terms, as part of a story about how we can know a lot about kinds that are not always easily observable (Boyd, 1988, 195-196). By analogy, he suggests that we should be optimistic that a naturalistically acceptable moral theory exists, and that it is consistent with us having a lot of moral knowledge.

Once we look at the details of Boyd's argument, we see that it is an argument that duelling intuitions about the Gödel/Schmidt example simply can't touch. In part that's because Boyd cares primarily about natural kind terms, not names. But more importantly it is because, as we noted in section 2, the only point that's at issue by the time Kripke raises the Gödel/Schmidt example is the *number* of descriptive names. Just looking at the arguments Kripke raises before that example gives us more than enough evidence to use in the kind of argument Boyd is making.

It would take us far beyond the length of a short reply to go through every philosophical use of the causal-historical theory that MMNS purport to refute in this much detail. But we think that the kind of response we've used here will frequently work. That is, we think few, if any, of the arguments they attack use the parts of the causal-historical theory that Kripke is defending with the Gödel/Schmidt example, and so even if that example fails, it wouldn't undermine those theories.

Bibliography

Adams, Douglas. 1980. *The Restaurant at the End of the Universe*. London: Pan Macmillan. Reprinted in **?**. References to Reprint.

Adams, Robert. 1974. "Theories of Actuality." *Noûs* 8:211–231.

Armstrong, D. M. 1978. *Universals and Scientific Realism*. Cambridge: Cambridge University Press.

—. 2000. "Black Swans: The Formative Influences in Australian Philosophy." In Berit Brogaard and Barry Smith (eds.), *Rationality and Irrationality*, 11–17. Kirchberg: Austrian Ludwig Wittgenstein Society.

Austin, J. L. 1962. *Sense and Sensibilia*. Oxford: Oxford University Press.

Ayer, Alfred. 1936. *Language, Truth and Logic.* London: Gollantz.

Bateman, Bradley. 1996. *Keynes's Uncertain Revolution.* Ann Arbor: University of Michigan Press.

Bays, Timothy. 2007. "The Problem with Charlie: Some Remarks on Putnam, Lewis and Williams." *Philosophical Review* 116:401–425, doi:10.1215/00318108-2007-003.

Bealer, George. 1998. "Intuition and the Autonomy of Philosophy." In DePaul and Ramsey (1998), 201–240.

Bennett, Jonathan. 1984. "Counterfactuals and Temporal Direction." *Philosophical Review* 93:57–91.

—. 2003. *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Bigelow, John. 1988. *The Reality of Numbers: A Physicalist's Philosophy of Mathematics*. Oxford: Oxford.

BIGELOW, JOHN, COLLINS, JOHN, and PARGETTER, ROBERT. 1993. "The Big Bad Bug: What are the Humean's Chances?" *The British Journal for the Philosophy of Science* 44:443–462, doi:10.1093/bjps/44.3.443.

Bochvar, D. A. 1939. "On a Three Valued Calculus and Its Application to the Analysis of Contradictories." *Matematicheskii Sbornik* 4:287–308.

Borel, Emile. 1924. "A propos d'un Traité de Probabilités." *Revue Philosophique* 98:321–336.

Boyd, Richard. 1988. "How to Be a Moral Realist." In Geoffrey Sayre-McCord (ed.), *Essays in Moral Realism*, 181–228. Ithaca: Cornell University Press.

Braddon-Mitchell, David and Nola, Robert. 1997. "Ramsification and Glymour's Counterexample." *Analysis* 57:167–169.

Bradford, Wylie and Harcourt, Geoff. 1997. "Definitions and Units." In Harcourt and Riach (1997), 107–131.

Brewer, Bill. 1999. *Perception and Reason*. Oxford: Oxford University Press.

Butterfield, Jeremy. 2006. "Against Pointillisme about Mechanics." *British Journal for the Philosophy of Science* 57:709–753.

Byrne, Alex. 1993. "Truth in Fiction - The Story Continued." *Australasian Journal of Philosophy* 71:24–35.

Cappelen, Herman. 2012. *Philosophy without Intuitions*. Oxford: Oxford University Press.

Carnap, Rudolf. 1950. *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Carruthers, Peter. 1990. *The Metaphysics of the Tractatus*. Cambridge: Cambridge University Press.

—. 2011. *The Opacity of Mind: an integrative theory of self-knowledge*. Oxford: Oxford University Press.

Chudnoff, Elijah. 2011. "What Should a Theory of Knowledge Do?" *Dialectica* 65:561–579, doi:10.1111/j.1746-8361.2011.01285.x.

Churchland, Paul. 1981. "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy* 78:67–90.

Coates, John. 1996. *The Claims of Common Sense*. Cambridge: Cambridge University Press.

—. 1997. "Keynes, Vague Concepts and Fuzzy Logic." In Harcourt and Riach (1997), 244–260.

Cohen, Stewart. 1986. "Knowledge and Context." *The Journal of Philosophy* 83:574–583.

Cummins, Robert. 1998. "Reflection on Reflective Equilibrium." In DePaul and Ramsey (1998), 113–128.

Currie, Gregory. 1990. *The Nature of Fiction*. Cambridge: Cambridge University Press.

—. 2002. "Desire in Imagination." In Gendler and Hawthorne (2002), 201–221.

Davies, Martin. 1981. *Meaning, Quantification, Necessity: Themes in Philosophical Logic*. London: Routledge.

Davis, John. 1994. *Keynes's Philosophical Development*. Cambridge: Cambridge University Press.

—. 1995. "Keynes' Later Philosophy." *History of Political Economy* 27:237–260.

DePaul, Michael and Ramsey, William (eds.). 1998. *Rethinking Intuition*. Lanham: Rowman & Littlefield.

DeRose, Keith. 1995. "Solving the Skeptical Problem." *Philosophical Review* 104:1–52.

—. 1996. "Knowledge, Assertion and Lotteries." *Philosophical Review* 74:568–79.

Deutsch, Max. 2009. "Experimental Philosophy and the Theory of Reference." *Mind and Language* 24:445–466.

Devitt, Michael. 2010. "Experimental Semantics." *Philosophy and Phenomenological Research* Forthcoming.

Devitt, Michael and Sterelny, Kim. 1987. *Language and Reality: An Introduction to the Philosophy of Language*. Cambridge, MA.: MIT Press.

Dummett, Michael. 1959. "Truth." *Proceedings of the Aristotelian Society* New Series 59:141–62.

Egan, Andy. 2007. "Epistemic Modals, Relativism and Assertion." *Philosophical Studies* 133:1–22.

Elga, Adam. 2000a. "Self-Locating Belief and the Sleeping Beauty Problem." *Analysis* 60:143–147.

—. 2000b. "Self-Locating Belief and the Sleeping Beauty Problem." *Analysis* 60:143–7.

—. 2004. "Defeating Dr. Evil with Self-Locating Belief." *Philosophy and Phenomenological Research* 69:383–396.

Field, Hartry. 1973. "Theory Change and the Indeterminacy of Reference." *Journal of Philosophy* 70:462–81.

Fine, Kit. 1975a. "Critical Notice of *Counterfactuals*." *Mind* 84:451–458.

—. 1975b. "Vagueness, Truth and Logic." *Synthese* 30:265–300.

Fodor, Jerry. 2000. *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.

Fodor, Jerry A. 1983. *The Modularity of Mind*. Cambridge, MA: MIT Press.

—. 1987. *Psychosemantics*. Cambridge, MA: MIT Press.

—. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

Fodor, Jerry A. and Lepore, Ernest. 1992. *Holism: A Shopper's Guide*. Cambridge: Blackwell.

Forrest, Peter. 1982. "Occam's Razor and Possible Worlds." *Monist* 65:456–464.

Forrest, Peter and Armstrong, D. M. 1984. "An Argument Against David Lewis' Theory of Possible Worlds." *Australasian Journal of Philosophy* 62:164–168.

Frankfurt, Harry. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68:5–20.

Friedman, Milton. 1953. "The Methodology of Positive Economics." In *Essays in Positive Economics*, 3–43. Chicago: University of Chicago Press.

Gendler, Tamar Szab□o. 2000. "The Puzzle of Imaginative Resistance." *Journal of Philosophy* 97:55–81.

Gendler, Tamar Szabó and Hawthorne, John (eds.). 2002. *Conceivability and Possibility*. Oxford: Oxford University Press.

Gladwell, Malcolm. 2005. *Blink: The Power of Thinking Without Thinking*. New York: Little, Brown.

Goldman, Alvin I. 1976. "Discrimination and Perceptual Knowledge." *The Journal of Philosophy* 73:771–791.

Goodman, Nelson. 1955. *Fact, Fiction and Forecast,*. Cambridge: Harvard University Press.

Gopnik, Alison. 2009. *The Philosophical Baby: What Children's Minds Tell Us About Truth, Love, and the Meaning of Life*. New York: Farrar, Straus and Giroux.

Grice, H. Paul. 1989. *Studies in the Way of Words*. Cambridge, MA.: Harvard University Press.

Haack, Susan. 1974. *Deviant Logic*. Chicago: University of Chicago Press.

Hájek, Alan. 2010. "David Lewis." In *The New Dictionary of Scientific Biography*. New York: Scribners.

Hall, Ned. 1994. "Correcting the Guide to Objective Chance." *Mind* 103:505–518.

Harcourt, G. C. and Riach, P. A. (eds.). 1997. *A 'Second Edition' of the General Theory*. London: Routledge.

Hare, R. M. 1951. *The Language of Morals*. Oxford: Oxford University Press.

Harman, Gilbert. 1973. *Thought*. Princeton: Princeton University Press.

—. 1986. *Change in View*. Cambridge, MA: Bradford.

Haslanger, Sally. 1994. "Humean Supervenience and Enduring Things." *Australasian Journal of Philosophy* 72:339–359, doi:10.1080/00048409412346141.

Hawthorne, John. 1990. "A Note on Languages and Language." *Australasian Journal of Philosophy* 68:116–118.

—. 2007. "Craziness and Metasemantics." *Philosophical Review* 116:427–440, doi:10.1215/00318108-2007-004.

Hetherington, Stephen. 2001. *Good Knowledge, Bad Knowledge: on two dogmas of epistemology*. Oxford: Oxford University Press.

Holton, Richard. 1997. "Some Telling Examples:Reply to Tsohatzidis." *Journal of Pragmatics* 28:625–8.

—. 2003. "David Lewis's Philosophy of Language." *Mind and Language* 18:286–295, doi:10.1111/1468-0017.00228.

Horowitz, Tamara. 1998. "Philosophical Intuitions and Psychological Theory." *Ethics* 108:367–85.

Horwich, Paul. 1999. *Meaning*. Oxford: Oxford University Press.

Humberstone, I. L. 1996. "Intrinsic/Extrinsic." *Synthese* 108:205–67.

Hume, David. 1757. "On the Standard of Taste." In *Essays: Moral, Political and Legal*, 227–249. Indianapolis: Liberty Press.

Humphreys, Paul and Fetzer, James (eds.). 1998. *The New Theory of Reference*. Dordrecht: Kluwer.

Ichikawa, Jonathan, Maitra, Ishani, and Weatherson, Brian. 2012. "In Defence of a Kripkean Dogma." *Philosophy and Phenomenological Research* 85:56–68, doi:10.1111/j.1933-1592.2010.00478.x.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Clarendon Press: Oxford.

Joyce, James. 1914/2000. *Dubliners*. Oxford: Oxford University Press.

—. 1922/1993. *Ulysses*. Oxford: Oxford University Press.

—. 1944/1963. *Stephen Hero*. New Directions: Norfolk, CT.

Joyce, James M. 1998. "A Non-Pragmatic Vindication of Probabilism." *Philosophy of Science* 65:575–603.

Kahneman, Daniel. 2011. *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.

Keefe, Rosanna. 2000. *Theories of Vagueness*. Cambridge: Cambridge University Press.

Keynes, John Maynard. 1909. "The Method of Index Numbers with Special Reference to the Measurement of General Exchange Value." In Keynes (1971-1989), 50–156.

—. 1921. *Treatise on Probability*. London: Macmillan.

—. 1931. "Review of *Foundations of Mathematics* by Frank Ramsey." *The New Statesman and Nation* 2:407;. Reprinted in (Keynes, 1971-1989, X 336-339).

—. 1934. *Draft of the General Theory*, 423–449. Volume XIII of Keynes (1971-1989).

—. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

—. 1937. "The General Theory of Employment." *Quarterly Journal of Economics* 51:209–223. Reprinted in (Keynes, 1971-1989, XIV 109-123), references to reprint.

—. 1938a. *Letter to Hugh Townshend dated 7 December*, 293–294. Volume 14 of Keynes (1971-1989).

—. 1938b. "My Early Beliefs." In Keynes (1971-1989), 433–451.

—. 1971-1989. *The Collected Writings of John Maynard Keynes*. London: Macmillan.

Kidd, John. 1988. "The Scandal of 'Ulysses'." *The New York Review of Books* 35:32–39.

Kieran, Matthew and Lopes, Dominic McIver (eds.). 2003. *Imagination, Philosophy and the Arts*. London. Routledge.

Kilkarni, Sanjeev and Harman, Gilbert. 2011. *An Elementary Introduction to Statistical Learning Theory*. Hoboken, NJ: Wiley.

Kim, Jaegwon. 1973. "Causes and Counterfactuals." *Journal of Philosophy* 70:570–572.

Klein, Gary A. 1999. *Sources of Power*. Cambridge, MA.: MIT Press.

Kripke, Saul. 1975. "Outline of a Theory of Truth." *Journal of Philosophy* 72:690–716.

—. 1980. *Naming and Necessity*. Cambridge: Harvard University Press.

—. 1982. *Wittgenstein on Rules and Private Language*. Oxford: Basil Blackwell.

Langton, Rae and Lewis, David. 1998. "Defining 'Intrinsic'." *Philosophy and Phenomenological Research* 58:333–345. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 116-132.

—. 2001. "Marshall and Parsons on 'Intrinsic'." *Philosophy and Phenomenological Research* 63:353–355.

Lewis, David. 1966. "An Argument for the Identity Theory." *Journal of Philosophy* 63:17–25. Reprinted with additions as **?**.

—. 1968. "Counterpart Theory and Quantified Modal Logic." *Journal of Philosophy* 65:113–126. Reprinted in *Philosophical Papers*, Volume I, pp. 26-39.

—. 1969a. *Convention: A Philosophical Study*. Cambridge: Harvard University Press.

—. 1969b. "Lucas against Mechanism." *Philosophy* 44:231–3. Reprinted in *Papers in Philosophical Logic*, pp. 166-169.

—. 1970a. "Anselm and Actuality." *Noûs* 4:175–188. Reprinted in *Philosophical Papers*, Volume I, pp. 10-20.

—. 1970b. "How to Define Theoretical Terms." *Journal of Philosophy* 67:427–446. Reprinted in *Philosophical Papers*, Volume I, pp. 78-95.

—. 1971. "Counterparts of Persons and Their Bodies." *Journal of Philosophy* 68:203–211. Reprinted in *Philosophical Papers*, Volume I, pp. 47-54.

—. 1972. "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy* 50:249–58. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 248-261.

—. 1973a. "Causation." *Journal of Philosophy* 70:556–567. Reprinted in *Philosophical Papers*, Volume II, pp. 159-172.

—. 1973b. *Counterfactuals*. Oxford: Blackwell Publishers.

—. 1974a. "Radical Interpretation." *Synthese* 27:331–344. Reprinted in *Philosophical Papers*, Volume I, pp. 108-118.

—. 1974b. "'Tensions." In Milton K. Munitz and Peter K. Unger (eds.), *Semantics and Philosophy*, 49–61. New York: New York University Press. Reprinted in *Philosophical Papers*, Volume I, pp. 250-260.

—. 1975a. "Adverbs of Quantification." In *Formal Semantics of Natural Language*, 3–15. Cambridge: Cambridge University Press. Reprinted in *Papers in Philosophical Logic*, pp. 5-20.

—. 1975b. "Languages and Language." In *Minnesota Studies in the Philosophy of Science*, volume 7, 3–35. Minneapolis: University of Minnesota Press. Reprinted in *Philosophical Papers*, Volume I, pp. 163-188.

—. 1976a. "The Paradoxes of Time Travel." *American Philosophical Quarterly* 13:145–152. Reprinted in *Philosophical Papers*, Volume II, pp. 67-80.

—. 1976b. "Probabilities of Conditionals and Conditional Probabilities." *Philosophical Review* 85:297–315. Reprinted in *Philosophical Papers*, Volume II, pp. 133-152.

—. 1978a. "Reply to McMichael." *Analysis* 38:85–86. Reprinted in *Papers in Ethics and Social Philosophy*, pp. 34-36.

—. 1978b. "Truth in Fiction." *American Philosophical Quarterly* 15:37–46. Reprinted in *Philosophical Papers*, Volume I, pp. 261-275.

—. 1979a. "Attitudes *De Dicto* and *De Se*." *Philosophical Review* 88:513–543. Reprinted in *Philosophical Papers*, Volume I, pp. 133-156.

—. 1979b. "Counterfactual Dependence and Time's Arrow." *Noûs* 13:455–476. Reprinted in *Philosophical Papers*, Volume II, pp. 32-52.

—. 1979c. "Lucas against Mechanism II." *Canadian Journal of Philosophy* 9:373–6. Reprinted in *Papers in Philosophical Logic*, pp. 170-173.

—. 1979d. "Prisoners' Dilemma is a Newcomb Problem." *Philosophy and Public Affairs* 8:235–240. Reprinted in *Philosophical Papers*, Volume II, pp. 299-304.

—. 1979e. "Scorekeeping in a Language Game." *Journal of Philosophical Logic* 8:339–359. Reprinted in *Philosophical Papers*, Volume I, pp. 233-249.

—. 1980a. "Mad Pain and Martian Pain." In Ned Block (ed.), *Readings in the Philosophy of Psychology*, volume I, 216–232. Cambridge: Harvard University Press. Reprinted in *Philosophical Papers*, Volume I, pp. 122-130.

—. 1980b. "A Subjectivist's Guide to Objective Chance." In *Studies in Inductive Logic and Probability*, volume 2, 83–132. Berkeley: University of California Press. Reprinted in *Philosophical Papers*, Volume II, pp. 83-113.

—. 1980c. "Veridical Hallucination and Prosthetic Vision." *Australasian Journal of Philosophy* 58:239–249, doi:10.1080/00048408012341251. Reprinted in *Philosophical Papers*, Volume II, pp. 273-286.

—. 1981a. "Are we Free to Break the Laws?" *Theoria* 47:113–121. Reprinted in *Philosophical Papers*, Volume II, pp. 291-298.

—. 1981b. "Causal Decision Theory." *Australasian Journal of Philosophy* 59:5–30. Reprinted in *Philosophical Papers*, Volume II, pp. 305-337.

—. 1981c. "What Puzzling Pierre Does Not Believe." *Australasian Journal of Philosophy* 59:283–289, doi:10.1080/00048408112340241. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 408-417.

—. 1982. "Logic for Equivocators." *Noûs* 16:431–441. Reprinted in *Papers in Philosophical Logic*, pp. 97-110.

—. 1983a. "Individuation by Acquaintance and by Stipulation." *Philosophical Review* 92:3–32. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 373-402.

—. 1983b. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61:343–377, doi:10.1080/00048408312341131. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 8-55.

—. 1983c. *Philosophical Papers*, volume I. Oxford: Oxford University Press.

—. 1984a. "Devil's Bargains and the Real World." In Douglas Maclean (ed.), *The Security Gamble: Deterrence in the Nuclear Age*, 141–154. Totowa, NJ: Rowman and Allenheld. Reprinted in *Papers in Ethics and Social Philosophy*, pp. 201-218.

—. 1984b. "Putnam's Paradox." *Australasian Journal of Philosophy* 62:221–236, doi:10.1080/00048408412340013. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 56-77.

—. 1986a. "Events." In *Philosophical Papers*, volume II, 241–269. Oxford: OUP.

—. 1986b. *On the Plurality of Worlds*. Oxford: Blackwell Publishers.

—. 1986c. *Philosophical Papers*, volume II. Oxford: Oxford University Press.

—. 1986d. "Probabilities of Conditionals and Conditional Probabilities II." *Philosophical Review* 95:581–589. Reprinted in *Papers in Philosophical Logic*, pp. 57-65.

—. 1988a. "Ayer's First Empiricist Criterion of Meaning: Why Does it Fail?" *Analysis* 48:1–3. Reprinted in *Papers in Philosophical Logic*, pp. 156-158.

—. 1988b. "Desire as Belief." *Mind* 97:323–32. Reprinted in *Papers in Ethics and Social Philosophy*, pp. 42-54.

—. 1988c. "The Trap's Dilemma." *Australasian Journal of Philosophy* 66:220–223. Reprinted in *Papers in Ethics and Social Philosophy*, pp. 95-100.

—. 1988d. "What Experience Teaches." *Proceedings of the Russellian Society* 13:29–57. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 262-290.

—. 1989a. "Academic Appointments: Why Ignore the Advantage of Being Right?" *Ormond Papers* 6. Reprinted in *Papers in Ethics and Social Philosophy*, pp. 187-200.

—. 1989b. "Dispositional Theories of Value." *Proceedings of the Aristotelian Society* Supplementary Volume 63:113–137. Reprinted in *Papers in Ethics and Social Philosophy*, pp. 68-94.

—. 1989c. "Mill and Milquetoast." *Australasian Journal of Philosophy* 67:152–171, doi:10.1080/00048408912343741. Reprinted in *Papers in Ethics and Social Philosophy*, pp. 159-186.

—. 1990. "Noneism or Allism?" *Mind* 99:23–31. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 152-163.

—. 1991. *Parts of Classes*. Oxford: Blackwell.

—. 1992. "Meaning without Use: Reply to Hawthorne." *Australasian Journal of Philosophy* 70:106–110, doi:10.1080/00048408112340093. Reprinted in *Papers in Ethics and Social Philosophy*, pp. 145-151.

—. 1993a. "Evil for Freedom's Sake?" *Philosophical Papers* 22:149–172. Reprinted in *Papers in Ethics and Social Philosophy*, pp. 101-127.

—. 1993b. "Mathematics is Megethology." *Philosophia Mathematica* 3:3–23. Reprinted in *Papers in Philosophical Logic*, pp. 203-230.

—. 1994a. "Humean Supervenience Debugged." *Mind* 103:473–490. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 224-247.

—. 1994b. "Reduction of Mind." In Samuel Guttenplan (ed.), *A Companion to the Philosophy of Mind*, 412–431. Oxford: Blackwell, doi:10.1017/CBO9780511625343.019. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 291-324.

—. 1995. "Should a Materialist Believe in Qualia?" *Australasian Journal of Philosophy* 73:140–44, doi:10.1080/00048409512346451. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 325-331.

—. 1996a. "Desire as Belief II." *Mind* 105:303–13. Reprinted in *Papers in Ethics and Social Philosophy*, pp. 55-67.

—. 1996b. "Elusive Knowledge." *Australasian Journal of Philosophy* 74:549–567, doi:10.1080/00048409612347521. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 418-446.

—. 1997a. "Do We Believe in Penal Substitution?" *Philosophical Papers* 26:203–209. Reprinted in *Papers in Ethics and Social Philosophy*, pp. 128-135.

—. 1997b. "Finkish Dispositions." *Philosophical Quarterly* 47:143–158. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 133-151.

—. 1997c. "Naming the Colours." *Australasian Journal of Philosophy* 75:325–42, doi:10.1080/00048409712347931. Reprinted in *Papers in Metaphysics and Epistemology*, pp. 332-358.

—. 1998. *Papers in Philosophical Logic*. Cambridge: Cambridge University Press.

—. 1999a. *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.

—. 1999b. "Why Conditionalize?" In *Papers in Metaphysics and Epistemology*, 403–407. Cambridge University Press. Originally written as a course handout in 1972.

—. 2000. *Papers in Ethics and Social Philosophy*. Cambridge: Cambridge University Press.

—. 2001a. "Redefining 'Intrinsic'." *Philosophy and Phenomenological Research* 63:381–398.

—. 2001b. "Sleeping Beauty: Reply to Elga." *Analysis* 61:171–176.

—. 2001c. "Truthmaking and Difference-Making." *Noûs* 35:602–615.

—. 2002a. "Tensing the Copula." *Mind* 111:1–14.

—. 2002b. "Tharp's Third Theorem." *Analysis* 62:95–97.

—. 2003. "Things qua Truthmakers." In Hallvard Lillehammer and Gonzalo Rodriguez-Pereyra (eds.), *Real Metaphysics: Essays in Honour of D. H. Mellor*, 25–38. London: Routledge.

—. 2004a. "Causation as Influence." In John Collins, Ned Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*, 75–106. Cambridge: MIT Press.

—. 2004b. "How Many Lives has Schrödinger's Cat?" *Australasian Journal of Philosophy* 82:3–22, doi:10.1080/713659799.

—. 2004c. "Void and Object." In John Collins, Ned Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*, 277–290. Cambridge: MIT Press.

—. 2007. "Divine Evil." In Louise Anthony (ed.), *Philosophers Without Gods*, 231–242. Oxford: Oxford University Press.

Lewis, David and Lewis, Stephanie. 1970. "Holes." *Australasian Journal of Philosophy* 48:206–212, doi:10.1080/00048407012341181. Reprinted in *Philosophical Papers*, Volume I, pp. 3-9.

Ludwig, Kirk. 2007. "The Epistemology of Thought Experiments." *Midwest Studies in Philosophy* 31:128–159.

Machery, Eduoard, Mallon, Ron, Nichols, Shaun, and Stich, Stephen. 2012. "If Folk Intuitions Vary, Then What?" *Philosophy and Phenomenological Research* , doi:10.1111/j.1933-1592.2011.00555.xForthcoming.

Maitra, Ishani. 2007. "How and Why to Be a Moderate Contextualist." In Gerhard Preyer and Georg Peter (eds.), *Context Sensitivity and Semantic Minimalism: New Essays on Semantics and Pragmatics*, 111–132. Oxford: Oxford University Press.

Mallon, Ron, Machery, Eduoard, Nichols, Shaun, and Stich, Stephen. 2009. "Against Arguments from Reference." *Philosophy and Phenomenological Research* 79:332–356.

Martí, Genoveva. 2009. "Against Semantic Multi-Culturalism." *Analysis* 69:42–48.

Matravers, Derek. 2003. "Fictional Assent and the (So-Called) "Puzzle of Imaginative Resistance"." In Kieran and Lopes (2003), 91–108.

Maudlin, Tim. 1994. *Quantum Non-Locality and Relativity: Metaphysical Intimations of Modern Physics*. Oxford: Blackwell.

—. 2007. *The Metaphysics Within Physics*. Oxford: Oxford University Press.

McDowell, John. 1996. *Mind and World*. Cambridge, MA: Harvard University Press.

Melia, Joseph. 1992. "A Note on Lewis's Ontology." *Analysis* 52:191–192.

Melia, Joseph and Saatsi, Juha. 2006. "Ramseyfication and Theoretical Content." *British Journal for the Philosophy of Science* 57:561–585.

Menzies, Peter. 1996. "Probabilistic Causation and the Pre-emption Problem." *Mind* 105:85–117.

Moggridge, Donald. 1992. *Maynard Keynes: An Economist's Biography*. London: Routledge.

Moran, Richard. 1995. "The Expression of Feeling in Imagination." *Philosophical Review* 103:75–106.

Nagel, Jennifer. 2007. "Epistemic Intuitions." *Philosophy Compass* 2:792–819, doi:10.1111/j.1747-9991.2007.00104.x.

—. 2013. "Intuitions and Experiments: A Defense of the Case Method in Epistemology." *Philosophy and Phenomenological Research* 85:495–527, doi:10.1111/j.1933-1592.2012.00634.x.

Nelkin, Dana. 2000. "The Lottery Paradox, Knowledge, and Rationality." *Philosophical Review* 109:373–409.

Nesbø, Jo. 2009. *The Redeemer*. London: Vintage Books.

Nolan, Daniel. 1996. "Recombination Unbound." *Philosophical Studies* 84.

—. 2005. *David Lewis*. Chesham: Acumen Publishing.

—. 2007. "Selfless Desires." *Philosophy and Phenomenological Research* 73:665–679.

O'Donnell, Rod. 1989. *Keynes: Philosophy, Economics and Politics*. London: Macmillan.

—. 1991. "Reply." In Rod O'Donnell (ed.), *Keynes as Philosopher-Economist*, 78–102. London: Macmillan.

—. 1997. "Keynes and Formalism." In Harcourt and Riach (1997), 131–165.

Plantinga, Alvin. 1974. *The Nature of Necessity*. Oxford: Oxford University Press.

Priest, Graham. 1999. "Sylvan's Box: A Short Story and Ten Morals." *Notre Dame Journal of Formal Logic.* 38:573–582.

Pryor, James. 2000. "The Sceptic and the Dogmatist." *Noûs* 34:517–549, doi:10.1111/0029-4624.00277.

Putnam, Hillary. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA.: MIT Press.

Ramsey, Frank. 1926. "Truth and Probability." In Ramsey (1990), 52–94.

—. 1929. "Probability and Partial Belief." In Ramsey (1990), 95–96.

—. 1931. *The Foundations of Mathematics and other Logical Essays*. London: Routledge.

—. 1990. *Philosophical Papers*. Cambridge: Cambridge University Press.

Ramsey, William. 1998. "Prototypes and Conceptual Analysis." In DePaul and Ramsey (1998), 161–177.

Robinson, Denis. 1989. "Matter, Motion and Humean Supervenience." *Australasian Journal of Philosophy* 67:394–409.

Rosch, Eleanor and Mervis, Carolyn. 1975. "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Science* 8:382–439.

Ruetsche, Laura. 2011. *Interpreting Quantum Theories*. Oxford: Oxford University Press.

Russell, Bertrand. 1923. "Vagueness." *Australasian Journal of Philosophy and Psychology* 1:84–92.

—. 1948. *Human Knowledge: Its Scope and Limits*. London: Allen and Unwin.

Ryle, Gilbert. 1949. *The Concept of Mind*. New York: Barnes and Noble.

—. 1954. *Dilemmas*. Cambridge: Cambridge University Press.

Sainsbury, Mark. 1995. "Vagueness, Ignorance and Margin for Error." *British Journal for the Philosophy of Science* 46:589–601.

Sartwell, Crispin. 1992. "Why Knowledge is Merely True Belief." *Journal of Philosophy* 89:167–180.

Schaffer, Jonathan. 2000. "Trumping Preemption." *Journal of Philosophy* 97:165–.

Schelling, Thomas. 1960. *The Strategy of Conflict*. Cambridge: Harvard University Press.

Schwarz, Wolfgang. 2006. "Lewisian Meaning without Naturalness." Draft, January 4, 2006. Downloaded from http://www.umsu.de/words/magnetism.pdf.

—. 2009. *David Lewis: Metaphysik und Analyse*. Paderborn: Mentis-Verlag.

Shope, Robert. 1983. *The Analysis of Knowledge*. Princeton: Princeton University Press.

Sider, Theodore. 1993. *Naturalness, Intrinsicality and Duplication*. Ph.D. thesis, University of Massachusetts - Amherst.

—. 2001a. "Criteria of Personal Identity and the Limits of Conceptual Analysis." *Philosophical Perspectives* 15:189–209, doi:10.1111/0029-4624.35.s15.10.

—. 2001b. *Four-Dimensionalism*. Oxford: Oxford University Press.

—. 2001c. "Maximality and Intrinsic Properties." *Philosophy and Phenomenological Research* 63:357–364.

—. 2002. "The Ersatz Pluriverse." *Journal of Philosophy* 99:279–315.

—. 2012. *Writing the Book of the World*. Oxford: Oxford University Press.

Skidelsky, Robert. 1983. *John Maynard Keynes. Vol. I: Hopes Betrayed, 1883-1920*. London: Macmillan.

—. 1992. *John Maynard Keynes. Vol. II: The Economist as Saviour, 1920-1937*. London: Macmillan.

Skinner, B. F. 1948. *Walden Two*. New York: Macmillan.

Smart, J. J. C. 1959. "Sensations and Brain Processes." *Philosophical Review* 68:141–156.

Smith, Michael. 2003. "Rational Capacities." In Sarah Stroud and Christine Tappolet (eds.), *Weakness of Will and Varities of Practical Irrationality*, 17–38. Oxford: Oxford University Press.

Soames, Scott. 1998a. "More Revisionism about Reference." In Humphreys and Fetzer (1998), 65–87.

—. 1998b. "Revisionism about Reference: A Reply to Smith." In Humphreys and Fetzer (1998), 13–35.

—. 2003. *Philosophical Analysis in the Twentieth Century*. Princeton: Princeton University Press.

Sorensen, Roy. 2001. *Vagueness and Contradiction*. Oxford: Oxford University Press.

Sosa, Ernest. 1998. "Minimal Intuition." In DePaul and Ramsey (1998), 257–269.

Stalnaker, Robert. 1984. *Inquiry*. Cambridge, MA: MIT Press.

—. 2004. "Lewis on Intentionality." *Australasian Journal of Philosophy* 82:199 – 212, doi:10.1080/713659796.

—. 2008. *Our Knowledge of the Internal World*. Oxford: Oxford University Press.

Stalnaker, Robert C. 1968. "A Theory of Conditionals." In Nicholas Rescher (ed.), *Studies in Logical Theory*, 98–112. Oxford: Blackwell.

—. 1976. "Possible Worlds." *Noûs* 10:65–75.

Sterelny, Kim. 2012. *The Evolved Apprentice: How Evolution Made Humans Unique*. Cambridge, MA.: Bradford.

Stich, Stephen. 1988. "Reflective Equilibrium, Analytic Epistemology and the Problem of Cognitive Diversity." *Synthese* 74:391–413.

—. 1992. "What is a Theory of Mental Representation?" *Mind* 101:243–63.

Stock, Kathleen. 2003. "The Tower of Goldbach and Other Impossible Tales." In Kieran and Lopes (2003), 107–124.

Strawson, Galen. 2000. "David Hume: Objects and Power." In Rupert Read and Kenneth A. Richman (eds.), *The New Hume Debate*, 31–51. London: Routledge.

Sugden, Robert. 2000. "Credible worlds: the status of theoretical models in economics." *Journal of Economic Methodology* 7:1–31, doi:10.1080/135017800362220.

—. 2009. "Credible Worlds, Capacities and Mechanisms." *Erkenntnis* 70:3–27, doi:10.1007/s10670-008-9134-x.

Teller, Paul. 1973. "Conditionalization and Observation." *Synthese* 26:218–258.

Tennant, Neil. 1992. *Autologic*. Edinburgh: Edinburgh University Press.

Thau, Michael. 1994. "Undermining and Admissibility." *Mind* 103:491–504.

Tribe, Kevin. 2002. "The Cambridge Economics Tripos 1903–55 and the Training of Economists." *The Manchester School* 68:222–248, doi:10.1111/1467-9957.00191.

Unger, Peter. 1996. *Living High and Letting Die*. Oxford: Oxford University Press.

van Fraassen, Bas. 1966. "Singular Terms, Truth–Value Gaps and Free Logic." *Journal of Philosophy* 66:481–95.

Walton, Kendall. 1990. *Mimesis as Make Believe*. Cambridge, MA: Harvard University Press.

—. 1994. "Morals in Fiction and Fictional Morality." *Aristotelian Society* 68(Supp):27–50.

Wang, Hao. 1987. *Reflections on Gôdel*. Cambridge, MA: MIT Press.

Warfield, Ted A. 2005. "Knowledge from Falsehood." *Philosophical Perspectives* 19:405–416, doi:10.1111/j.1520-8583.2005.00067.x.

Warnock, G. J. 1989. *J. L. Austin*. London: Routledge.

Weatherson, Brian. 2003a. "Many Many Problems." *Philosophical Quarterly* 53:481–501.

—. 2003b. "What Good Are Counterexamples?" *Philosophical Studies* 115:1–31, doi:10.1023/A:1024961917413.

—. 2004. "Luminous Margins." *Australasian Journal of Philosophy* 82:373 – 383.

—. 2005. "Scepticism, Rationalism and Externalism." *Oxford Studies in Epistemology* 1:311–331.

—. 2006. "The Asymmetric Magnets Problem." *Philosophical Perspectives* 20:479–492.

—. 2007. "The Bayesian and the Dogmatist." *Proceedings of the Aristotelian Society* 107:169–185, doi:10.1111/j.1467-9264.2007.00217.x.

—. 2009. "David Lewis." In Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*.

—. 2010. "Vagueness as Indeterminacy." In Richard Dietz and Sebastiano Moruzzi (eds.), *Cuts and Clouds: Vagueness, its Nature and its Logic*, 77–90. Oxford: Oxford University Press.

Weinberg, Jonathan, Stich, Stephen, and Nichols, Shaun. 2001. "Normativity and Epistemic Intuitions." *Philosophical Topics* 29:429–460.

Wettstein, Howard. 2004. *The Magic Prism*. Oxford: Oxford University Press.

Whewell, William. 1840. *The Philosophy of the Inductive Sciences, Founded Upon Their History*. London: John W. Parker.

Williams, J. Robert G. 2007. "Eligibility and Inscrutability." *Philosophical Review* 116:361–399, doi:10.1215/00318108-2007-002.

Williamson, Timothy. 1994. *Vagueness*. Routledge.

—. 2000. *Knowledge and its Limits*. Oxford University Press.

—. 2007. *The Philosophy of Philosophy*. Blackwell Pub. Ltd.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. London: Macmillan.

—. 1956. *Remarks on the Foundations of Mathematics*. New York: Macmillan.

Yablo, Stephen. 2002. "Coulda, Woulda, Shoulda." In Gendler and Hawthorne (2002), 441–492.