Intuitions and Conceptual Analysis

Week 7: Dealing with Counterexamples

1. Theory

Last week we looked at some cases where intuitions go wrong, and suggested that these all have something in common. In particular, I suggested that the problem in these cases is that the intuitions in question are in conflict with something loosely identified as the best theory. This week I want to start by saying more about makes a theory the best theory. (This is an elaboration of what I say in section 2 of the paper.) The order is changed from that paper because what is most helpful for polemical purposes is not necessarily the clearest!

Agreement with Theoretical Intuitions

As well as intuitions about the particular cases which are and aren't *F*s, we have intuitions about the theoretical role that *F* should play, about the inter-connections between *F*s, and so forth. One thing we seem to have quite a few intuitions about, for almost any interesting *F*, is the class of properties that *F* supervenes on. These intuitions may not be obvious, but it does seem that we have intuitions that two objects couldn't differ as to whether they were each *F* without differing in terms of some other property. For instance, I think intuition says that it is impossible for there to be two people with identical histories such that it is reasonable for one to believe *p* and not reasonable for the other to believe that *p*. (Polemical digression: if this is right then epistemic externalism has a very counter-intuitive consequence.) In other words, the reasonableness of an agent's belief supervenes on that agent's history.

This seems to be an instance of a very important theoretical intuition: normative concepts are, in some sense, non-arbitrary. I think this is the big message underlying the Horowitz paper we will look at in the second half of the class. The best arguments for radical utilitarianism, particularly in the form argued for by Peter Singer, turn on our intuition that moral concepts shouldn't be applied arbitrarily, and Singer's masterful presentation of the case that species membership is an arbitrary criterion. Now if there is nothing similar in the case of knowledge, that would explain why we take intuitions about possible cases more seriously in theory of knowledge (or particularly theory of causation) than in ethics.

It seems there is some evidence that this is not the case. As Jackson mentions, we have intuitions about which kinds of cases are saliently similar in theory of knowledge. Indeed, these intuitions are shared widely enough to do work in arguments that aim to change minds about Gettier-style cases. It isn't too hard to show that these

intuitions about what kinds of properties make cases the same or different can be rephrased as supervenience intuitions. And supervenience intuitions are theoretical intuitions *par excellance*.

Theoretical Significance

There is some evidence that we want our conceptual terms to refer to theoretically significant classes of entities. We might think this is what should be taken away from the cases of whales, Mars and glass we discussed last week. That is, there is some evidence that Slote's Theory of Important Criteria is right.

I do believe that this is a constraint on analyses, but I should note that there is an important lacunae in the argument. In all the cases that Slote discusses, and which we discussed following Slote, we might not need to rely on such a constraint to get the right analysis. It might be that we have a *general* intuition that the terms refer to classes that are scientifically, or theoretically, significant. I think, but I'm not sure, that is what Jackson takes to be the lesson from the Slote cases.

One more thing on these cases. Sometimes it is suggested (particularly by Jackson's "Reference and Description Revisited") that the Slote cases are really cases of ambiguity. The suggestion is that 'solid' has two meanings, call them a folk meaning and a scientific meaning, and on the first it is true that glass is a solid, but on the second it isn't true. (I have heard conflicting reports on whether glass is a liquid or not, but it certainly isn't a solid in the scientific sense.) There are crucial data points that the ambiguity theory will, I think, struggle to explain. The ambiguity theory is that when we're in a context that the folk usage is appropriate, it can be right to say *Glass is a solid* but when we are in the science classroom, this locution becomes unacceptable. Even if this is right, it doesn't help the ambiguity claim much. We still want to know why the scientific usage has normative force outside the science classroom, but the folk usage has ***no*** normative force inside the science classroom. Further, we should have a theory as to why we are so happy to accept the scientific usage, and for that matter why we think we are learning a new fact, not a new word, when we accept it. But the clincher, I think, is that we can properly appeal to the scientific usage in settling debates in 'folk' contexts, but we cannot appeal to the folk usage to settle debates in scientific contexts.

Simplicity

My reasons for including this criteria turn mostly on moderately technical issues in the philosophy of language, one of which we'll talk about next week, so I don't want to spend too much time on it now. Having a simplicity criteria in our theory of meaning solves two rather pressing problems in semantics. The first problem is what has become known as the Kripkenstein problem. How can it be that meaning is determined by our dispositions to use terms when we think that the meaning of a word determines how it will apply in infinitely many cases, but we only have finitely

many dispositions? (I've never been convinced that we do have only finitely many dispositions; sure we are only of finite size, but the jury is still out on whether we can (I mean really can) take infinitely many different states. My thermostat is finite, but whether it can be in a finite or an infinite number of states depends on some unresolved questions in quantum physics.) Anyway, if we take the meaning of a conceptual term to be the simplest possible meaning consistent with our dispositions to use the term, the problem seems to vanish in the appropriate way. This, essentially, is what Lewis suggests at the end of "New Work".

The other problem is related to issues in pragmatics. Just to sketch the problem, in the 1950s some simple analyses of terms came under attack because they made apparently false predictions about how the terms would be used in simple cases. So it was claimed that 'knowledge' can't entail 'belief', because in paradigm cases of knowledge, such as my knowledge that I have two hands, it would be odd to say that I believed that I have two hands. Grice showed that, when supplemented by some fairly mundane principles governing conversational cooperativeness, the traditional theories did make correct predictions. It would be odd to say that I have two hands, not because that would be a *false* utterance, but because it wouldn't be a particularly helpful contribution to any ordinary helpful conversation. And the rule governing assertion isn't just *Say what's true*, but *Say what's true and helpful*.

So this was the start of a beautiful project, really one of the great achievements of philosophy since the moderns. (Does that make it post-modern?) But there's a gap in the argument. All Grice gave was a relative consistency proof: the traditional analyses are *consistent* with the data being as they actually are. But using similar arguments it can be shown that all sorts of non-traditional theories are also consistent with the data. In particular the kinds of alternatives to traditional analyses that were promoted in the 1950s are also consistent with the data, and don't do noticeably worse on the two tests we've mentioned so far. (They do better, though it isn't clear how much better, on the test we haven't mentioned yet.) So why can Grice's arguments be used to win the game for traditionalists, as opposed to fight out a stalemate? Grice says little about this, and what he says is patently wrong. I think the only possible answer is that simplicity is a criteria on analyses, and the traditional analyses are simpler than their 'ordinary language' rivals. (*Digression*: If simplicity is allowed as a criteria it also solves Quine's problems about the indeterminacy of translation. In fact I think Quine's problem is the *same* problem as the problem of choosing between traditional and ordinary language analyses. But this is a discussion for a philosophy of language class. *End of Digression*.)

Fewest Counterexamples

What I call the 'radical' position in my paper is the view that theory is under no obligation to agree with intuitions about possible cases. This seems quite implausible to me, although not all will agree. One tradition is theory of knowledge, the one I spend the most time discussing, takes intuitions about possible cases to of central importance.

Another tradition, a relatively famous one, stresses theoretical intuitions to the exclusion of almost everything else. There is no argument from possible cases for Cartesian scepticism! If we just look at theoretical principles, we might (*might*) be led to think that knowledge requires a kind of certainty we couldn't ever obtain. But this is just an incredible position. We all know all too well so many facts about the world. It would be nice, in a way, to not know about all the wars that have gone on in recent times. (It would be nicer still if some friendly metaphysician of an eliminitavist bent could show there are no wars!) But this does not describe our situation. Just sometimes, Moorean arguments are sound, but the radical says they are never sound, so the radical is wrong. The interesting question, one we will spend some time on, is how to assess the relative importance of different examples. I'll suggest three criteria that are crucial, and one that almost certainly is not.

First, we are under more of an obligation to capture strong intuitions than weak ones. If we have a weak intuition that *a* is *F*, but theory says it is not an *F*, this is a minor problem for the theory, not a major problem.

Secondly, the *breadth* of the example matters. An example that applies in a wider range of cases is more important than one that only applies in very specific cases, I think.

Thirdly, our *familiarity* with the examples matters. Intuitions are unreliable about distant cases, and we should distrust anyone who claims to have strong intuitions about highly unfamiliar examples. What would count as a cause if there was no time and all matter was homogenous? Well I don't know, and I suspect no one else really does either.

What isn't relevant is whether the examples are *actual* or not. It could hardly save the JTB theory if we did an empirical investigation and found there were no Gettier cases in the actual world. We think our concepts apply to all sorts of merely possible cases, and as long as these are within the bounds of familiarity, theory should reflect these applications.

Three Arguments for This Theory

What I stressed last week was that a theory looking much like this deals best with cases where we agree that our former intuitions are mistaken. In all those cases, I hope, the mistaken intuition is in conflict with the 'best theory', as described here. Since I wrote the tests on good theories so as to deal with these cases, I have some confidence that this is so! So the first argument for the theory is that it correctly predicts that a variety of intuitions will be mistaken. I am here relying on a further intuition, that our earlier intuitions were mistaken, but this seems unavoidable.

The second argument for the theory is that it explains what happens in cases where we have differences of opinion over the intuitions. When my intuition is that *a* is an *F*, and yours isn't, sometimes the right response is to say what Jackson says, that we mean different things by *a* or *F*. But this isn't always what we do.

To take a concrete example, one that Jackson mentions, consider disputes over the nature of entailment. Few people have the intuition that a contradiction entails everything (*ex falso quadlibet*). Some people, including me, come to believe it after reviewing the evidence. But many people do not come to believe this, despite the existence of excellent arguments for it being true. One such argument, relying largely entirely on our theoretical intuitions about the role of entailment, is as follows:

(1) A entails A or B

(2) (A or B), not-A entails B

(3) Whenever (F entails C) and (C, D entails E) then (F, D entails E)

The last premise is a little complicated; it says that transitivity of entailment holds even when we start adding side-premises. Once this is seen, the intuition that entailment is transitive does the rest. Some logics do not have this feature, but it is not clear they are aiming to capture entailment in the traditional sense. (If you thought 'entails' means 'inductively implies', then you would reject transitivity of entailment, for reasonably well known reasons. But why would you ever think *that* to begin with?) Anyway, from these premises we get A, not-A entails B, which most everyone agrees implies that (A and not-A) entails B.

It is hard to deny (1) or (2), but they imply this very odd result that a contradiction entails everything. One possible response here would be to say that intuition shows (3) is wrong; after all intuition endorses (1) and (2), but not *ex falso*, so by *modus tollens*, (3) is false. It will be hard to give a *theory* of entailment without (3), but maybe what we should conclude is that intuition shows entailment is *sui generis*. As it turns out, this is not what logicians typically do. Rather, if they are to dispose of *ex falso*, they normally bite the bullet and deny that (2) is true. The trick then is to come up with a theory according to which entailment is a (relatively) simple property which has many of the general properties we pre-theoretically attribute to it, such as variants of transitivity like (3). If the aim of conceptual analysis, in this case the analysis of entailment, was just to capture the intuition, this project seems thoroughly misguided. But if I'm right, this is just how we should go about analysis. If there is a decent theory of entailment which does better on my criteria than the traditional theory, that theory is right. And if there are several competing theories which do equally well on my criteria, then there is no unique concept which 'entails' picks out. This is the position which seems to be becoming known as 'logical pluralism', and which if I'm right here is, at the very least, a rather viable option.

Sometimes we agree that a dispute over intuitions about particular cases can be resolved by appeal to theory. We agree, that is, to settle the question of whether *a* is *F*, about which intuitions divide, by waiting for the best theory to come in and seeing what verdict it gives. If we didn't agree to this, why would we spend so much time

on systematisation, time that could be better spent on the important task of listing the intuitions about every possible case, the only work that analysts need to do?

The final reason for liking my theory of analysis is that it fits so well with the best account of meaning for theoretical terms. Or more precisely, it fits with the 'theory' which lists the best solutions to the Kripkenstein and Grice paradoxes. Of course the reason it fits so well is because it was designed to incorporate those solutions.

Digression on Philosophy of Science

I think these arguments show that our intuitions about Gettier cases are not fatal to a justified true belief (JTB) analysis of knowledge. An analysis of knowledge is allowed to disagree with intuitions about particular cases, though remember I have argued that this disagreement is a cost. Anyway, for all we know at the end of inquiry it will turn out that the intuitions the best theory reforms are those about the Gettier cases, so the JTB theory is consistent with our having these intuitions.

What I don't stress in the paper, partially because it is so obviously true, is that the right thing to do when faced with our intuitions in Gettier-like cases is to do exactly what people actually did. That is, the right thing to do is to try to find a theory which is 'best' by my criteria. Now it turned out, surprisingly, that little headway could be made here. We could certainly reduce the range of the counterexamples by decreasing the simplicity of the theory, and at some great stages (particularly when Nozick's counterfactual theory was developed) we could reduce the intuitive force of the strongest counterexamples. Nozick's theory openly rejected closure of knowledge under ideal rationality, which struck me as a major theoretical shortcoming, but from my read of the literature a substantial number of people thought its false predictions in esoteric cases was a bigger problem.

There is an interesting comparison between the method I am using and some suggestions which have been made about the appropriate method for theory change in science. After Kuhn's *Structure of Scientific Revolutions*, many started to think that there were no rational criteria on theory change. Every theory could, in principle, deal with every anomaly by making small changes, or somehow impugning the data. Scientists didn't change theories when they thought the current theory couldn't explain the data; they changed theories when the current theory stopped being enough fun to work with. Imre Lakatos's theory of research programs, outlined in his *Methodology of Scientific Research Programs*, tried to account for some of the same historical data that interested Kuhn, without being as rationalist as the Kuhnians. On Lakatos's theory, a research program has a 'hard core', a set of views which distinguish the program, and a 'protective belt' of auxiliary assumptions and conjectures which are used to derive predictions about the world. For familiar reasons, no single data point can be used to show the hard core is mistaken; any mistaken prediction can be attributed to a mistake in the protective belt. A research program is replaced then, not when the hard core is shown to be wrong, but when (a) the changes needed to the protective belt to accommodate the

data start being too *ad hoc* and (b) the theory ceases to make interesting predictions. Importantly, the standards which must be met by a research program to stay alive are not absolute; rather, a research program simply must to better by these criteria than its rivals.

This all sounds pretty much like my theory in many respects. An analysis is not shown to be false just because it disagrees with one data, or intuition about a possible case. These disagreements with intuition are a cost, but if they can be explained away in a not too *ad hoc* manner, then we can live with them. What does sound the death knell for an analysis is the emergence of a rival which has all the qualities of the old analysis with none of the defects. There is, however, an interesting difference. I think that when anomalies start to emerge, the right response is often to look for a new analysis. It is only *after* we have failed to find this new analysis that it might be worth looking back at the old, discarded theory. On Lakatos's view of science, the right way to deal with a scientific anomaly is to see if it can be accommodated without a change to the research program. So there are strong analogies between my approach to choosing between analyses and Lakatos's theory of how to choose between research programs. However the analogy does not extend to views about what should be done in the face of anomalies.

2. The Infallibility of Meaning Intuitions

See section four of the paper

3. How to Beat a Counterexample

- Deny the Intuitions
- Intuitions are inconsistent; this is least plausible member
- Near Enough is Good Enough
- Not Much of an F
- False Implicit Theory
- Mistaken Identity
- Guilt by Association
- Pragmatic Semantic Confusion

I want to talk for a bit about what moves we can make to defend a theory when faced with a putative counterexample. For simplicity, I'll be assuming we're defend the theory that all and only *F*s are *G*s. The following batch of moves seem, in principle, to be available. (One interesting exercise is to see how many of these Lewis makes in "Causation as Influence", or even better, to discern other moves I have failed to notice.)

Near Enough is Good Enough

The motivation for this move is the following passage from **Lewis**.

> Maybe nothing could perfectly deserve the name "sensation" unless it were infallibly introspectible; or the name "simultaneity" unless it were a frame-independent equivalence relation; or the name "value" unless it couldn't possibly fail to attract anyone who was well acquainted with it. If so, then there are no perfect deservers of these names to be had. But it would be silly to loose our Moorings and deny that there existed any such things as sensations, simultaneity and values. In each case, an imperfect candidate may deserve the name quite well enough (246).

The fact that the property picked out by *G* isn't a perfect deserver of the name "*F*" is no evidence that there is a distinction to be had between the *F*s and the *G*s; it might be that there is nothing which could be a perfect deserver of that name. This is very similar to the argument in my paper, so I won't spend much time on this point, other than to rehearse a quick summary of the arguments.

First, when there is no perfect deserver, an imperfect deserver will usually do. Second, there is reason to think that the kind of cases we discuss on philosophy are cases where there will be no perfect deservers. The reason is that if there were a perfect deserver, it would have been discovered long ago, and the issue would have ceased to be a live one. Finally, the kind of counterexamples we usually discuss in philosophy are, by the nature of the subject, liable to be the kind of extreme cases where imperfect deservers go imperfect. If *G* imperfectly deserves the name "*F*", there will have to be some cases where intuition says that *a* is *F* and *a* is *G* differ in truth value. That's just what it is to be imperfect. It is better, *ceteris paribus*, for these cases to be extreme cases which are hard to think up and on which intuitions are not always unified or clear. In other words, once you've accepted that your analysis is imperfect, the last thing you should be worried about is a philosopher's counterexample.

Inconsistent Intuitions

Clearly if the intuitions are inconsistent, they aren't all true. And since it is no cost to be at odds with intuitions which are false, showing this may escape a counterexample. This seems to be the strategy that Unger was pursuing in the passages we looked at.

As it stands, this move is a little quick. That the intuitions here are inconsistent just shows that one of them must be false, not that all of them are. (Indeed, it will sometimes show that one of them must be true.) What we must show is that the intuition that we want to drop is the false one in the set. Now sometimes this will be quite easy: if the folk are inclined to drop the intuition we want them to drop when presented with the inconsistency, this seems to be sufficient to seal the case.

Even still, this move is often misapplied. The criteria can't be that under any old presentation of the possible inconsistency the folk choose to resolve it by dropping the intuition which we want them to drop. Just as first-order intuitions can vary with the presentation of the material, so can second-order intuitions, or intuitions about which intuitions are stronger than others. Given the esoteric nature of the case, the possible influence of the interrogator on subject at this point is particularly large.

And this says nothing about the really hard case, the case where the folk are not sure what to do when presented with the inconsistency. (Compare what happens when the folk are presented with the semantic paradoxes.) I suspect the methodological moral is that this defence may fail unless it is quite clear that the intuition we want people to drop is the one which is least strongly held.

There may be another move which can be made to save the defence, though it isn't often appealed to, especially around here. It might be argued that the fact of the inconsistency shows that the folk are unreliable about these matters, and we all know that we shouldn't trust unreliable sources. The problem is that this casts doubt not only on the counterexample, but on our 'evidence' that All and only *F*s are *G*s, which will usually be little more than its strong intuitive plausibility.

Scalar and Absolute Predicates

This is a move Lewis makes at a couple of crucial points in the causation paper. On Lewis's theory the presence of any object in the near-ish vicinity is a cause of every event, because of the gravitational and electro-magnetic forces objects bring with them. This is very odd, since it makes almost every object a cause of almost every event, since those forces tend to dissipate over rather large distances. Lewis's response is to accept that this is, in a sense true: the spatio-temporal location of any particular object is a cause of any particular event. But he denies that it is much of a cause. What we thought was an absolute, or on/off predicate, turns out to be essentially a scalar predicate. These weak attractive forces are very small causes of events; something is properly called a cause if it a sufficiently large cause. See page 15 of the paper for a discussion of this point, replete with some questionable analogies to similar moves regarding quantifiers.

The general point is that if there is a confusion between scalar and absolute uses of a predicate, the folk may well confuse something not being much of an *F* for it not being an *F*.

False Implicit Theory

Jackson says at a couple of points that we can give less weight to an intuition about a possible case if we can show that the 'intuition' isn't basic, but is derived from some more general intuition. This seems clearly wrong in general; why not say general intuitions are more likely to be right? But if we can show that the general intuition is wrong, and

that the only reason for the strong view about the possible case this view is forced by the general intuition, we seem to have won the game.

This may be the case in debates about personal identity across time and worlds. If someone has the intuition that the property 'being part of the same person' is intrinsic to a pair of stages, they will have all sorts of odd intuitions about hard cases in the identity literature. Since the intrinsicness intuition is (a) plausible and (b) false, recognising this could lead to placing a more appropriate weight on the folk intuitions.

This is less of a risk, but I guess many folk give some credence to the idea that it is impossible to have beliefs and desires without having conscious states. I don't know how plausible this is; it certainly sounds false to me. Anyway, perhaps this theory is behind the anti-functionalist intuitions in Chinese room and nation cases in philosophy of mind.

Mistaken Identity

This is to some extent a variant on the previous defence. Sometimes it is possible to show that an implicit theory is plausible but false by showing that there is a true theory with which it is easily confused. For instance, utilitarians deny that it is always wrong to execute innocents. But they agree that it is always wrong *ceteris paribus* to execute innocents, and that it is almost always wrong in practice to execute innocents, and so on. This can become important in metaphysics when there is the possibility for subtle confusions between intuitions of metaphysical possibility and intuitions of epistemic possibility.

For example, it seems possible that I could retain my identity while losing all the properties that are normally taken to be constitutive of my identity over time. There doesn't seem to be a contradiction in the story that runs: "Brian woke up one morning with an entirely new body and no memories of his former life. Had he been able to remember that he was on the run from the Mafia, he would have been quite pleased with this evasive technique." If personal identity over time goes by psychological continuity, or physical continuity, or some combination of the two, this story is incoherent. So why don't we take this clear possibility to refute those theories of identity. One possible out is to say that the reason this story seems possible *simpliciter* is that a Cartesian theory of identity, that identity means preservation of soul-stuff, is epistemically possible, and if that theory is true then the story I told is metaphysically possible.

Two More Moves

Next week we will try to fit in around the presentations two more moves which can be made here. The first is the move Horowitz makes in her paper in **Intuition**, of arguing that the intuition which is being relied upon is the very

same intuition which leads to clearly mistaken views in other areas. The second relies on Grice's important work on the relationship between what we can say, and what is true.

So the reading is the Horowitz paper, the two chapters from Grice which are in the filing cabinet, and the Lewis paper which is in the filing cabinet. In the Lewis paper, pay particular notice to the moves made in the last section, "Causation by Absences."