

Contra Williamson on Counterfactuals and Thought Experiments

Sharon Berry

August 30, 2009

1 Introduction

In Philosophy of Philosophy, Williamson considers how to represent the logical form of arguments by thought experiment, like the one involved in Gettier's famous argument about how one can have justified true belief without knowledge. Are we reasoning like this?

- (1) possibly, there is someone whose actual situation satisfies the stipulations made by Gettier
- (2) necessarily: everyone in the whose actual situation satisfies the stipulations made by Gettier counts as having JTB but not knowledge
- (conclusion) Possibly there is someone who has justified true belief but not knowledge.

This would be a valid argument, but, Williamson points out, it's not a good formalization of what we are doing. The problem is that Gettier's argument doesn't stipulate every single fact about Jones' situation, and it seem like there could be some scenarios under which all these stipulations are true, but Jones still doesn't count as having knowledge. As Williamson says, when evaluating thought experiments, we don't "worry about whether our verdicts would hold

even if mad scientists were interfering with the subject's brain processes in various ways; these possibilities do not normally occur to us when we assess Gettier counterexamples. Similarly, when moral philosophers assess imaginary examples, one can almost always fill out the case with unintended but morally relevant additions that would reverse the verdict. Any humanly compiled list of such interfering factors is likely to be incomplete".

Since we (presumably) don't want to represent Gettier's intuitively good argument, as having a strong and possibly false premise like (2), we need to look for a different formalization. And, here is Williamson's proposal, instead of using a claim about the everyone in all possible worlds, like (2), we should restrict our attention to the closest possible worlds, and use a claim like (2*) instead. 2* makes the weaker claim that the closest possible world in which someone satisfies the gettier stipulations, is one in which everyone who satisfies the gettier stipulations counts as having knowledge. Thus, we don't have to worry about mad scientist scenarios, when evaluating Gettier's claim, since these presumably take place very far away. So (he claims) a better formalization is:

- (1) possibly, there is someone whose actual situation satisfies the stipulations made by Gettier
- (2*) everyone in the closest possible world where someone actually satisfies the stipulations made by Gettier, counts as having JTB but not knowledge
- (conclusion) Possibly, there is someone who has justified true belief but not knowledge.

Now, does this fix the problem? Well, it rules out some scenarios, but (as Williamson himself points out) it still might be the case that the actual world contains someone who satisfies Gettier's stipulations but (because of other weird features of their case) doesn't count as having knowledge. I mean, here's an ex-

ample of the kind of moral thought experiment, which Williamson mentions in the quote above. Polemarchus says that justice is speaking the truth, and paying your debts. Socrates proposes a counter example: Suppose someone left a “deposit of arms” with you and then asked for it back while “not in his right mind”. In this situation, it would not be just to repay the debt. Now as Williamson suggests, surely there are some possible scenarios in which it *would* be just to return the arms to a madman. For example, it would presumably be just to return the weapon, if you had good evidence that the madman was already armed (so having extra weapons would make no difference). And, similarly, it would seem to be just to return the weapon if you had good evidence that the madman wouldn’t be violent if they got their weapon back, but would fly into a rage and kill you if they didn’t, or if giving them the weapon allowed other innocent people whom they would certainly have killed to escape.

The problem is, that not only are there possible scenarios like this, but in the long history of the world it would be surprising if there weren’t at least one actual case. And, if that’s true, then Williamson’s revised interpretation of Socrates’ argument will still have a false premise. For, consider the claim: if there were to be someone in the situation Socrates stipulates, then every one in that situation would be acting unjustly. Presumably, someone, at some point in the history of the world, has or will return a knife to a madman. So Williamson’s conditional requires that everyone who has returned a weapon to a madman acted unjustly in doing so. But is that true? If anything, I’m almost certain that situation of the kind mentioned above, (with a madman that’s already armed, or needs to be distracted in order to give his hostages time to escape) has already happened, or will happen at least once in the future.

Thus, it just looks like Williamson’s appeal to counterfactuals doesn’t even solve the problem which he proposed it to solve. Williamson admits the conflict

with intuition, but seems to think that we are better off. But this is not at all clear, is Socrates' counterexample really any less intuitively good than Gettier's? If not, it's hard to see what argument Williamson has against the initial, merely modal, interpretation of thought experiments, that does not apply, with equal strength, to his own position.

Now, Williamson is quite aware of this objection, and has a whole section on it. But, (so far as I can tell), all he does in this section is a) argue against various alternative ways of weakening (2) and b) propose an error theory about why intuition pulls against his theory in cases like the Socrates one above. The error theory, is that people claim that thought-experiments with false premises are actually correct, because they now see how they could fix them, and they don't want to admit that they were initially wrong. But this doesn't seem like a very good error theory. Firstly, it doesn't explain why I should think that Socrates' thought experiment argument was a good one. I might dislike admitting that I'm wrong, but that doesn't explain why I should be hesitant to say that (considering the empirical facts mentioned above), Plato was wrong. If anything, it would be rather flattering to my vanity to think I had saw something Plato missed. Similarly, note that in this case, I don't see how to 'fix' the argument (by adding further stipulations that make all actual instances true), and yet I still have the intuition that it's successful.

So, he never actually says why moving to his counterfactual proposal is in any way better off than the original, non-counterfactual version above. At best, he's shown that all the alternatives are equally bad. But he hasn't even done that. For one thing, he mentions that Gettier describes his case as a fiction about Smith and Jones and that "one could attempt an analysis of thought experiments that took their status as fictions more seriously, their relevance to fictional claims such as 1 is more easily understood in [a] more literal-minded

way”, and then never comes back to this possibility, even when all the literal accounts, (including his own) seem to suffer from problems as or worse than the Socrates problem above. Another (related possibility) would be to invoke ceterus paribus clauses. Perhaps the relevant version of two is something like: ceterus paribus, returning a knife to a madman is unjust. In this case we will need to tweak (1) a bit to get a valid argument, leaving something like.

- 1** Possibly: someone returns a knife to a madman in a situation where the ceterus are indeed paribus.
- 2** If someone returns a knife to a madman in a situation where all other things are being equal, they are repaying their debts but acting unjustly. (A third possibility along these lines might be to appeal to generics)

Insofar as Williamson’s argument for formalizing reasoning by thought experiments at all, is it must be possible to “capture what’s rational” in the thinking which leads a reader of the Gettier paper up to the conclusion that the JTB account of knowledge fails, you might think that Williamson would be equally committed to giving an account that captures what’s rational in reasoning about scientific reasoning involving ceterus paribus clauses, or literary debates about what’s true in a given fiction (certainly the participants in such debates about e.g. who the author of the poem in *Pale Fire* is, or whether the last few days in *Lolita* were a hallucination, seem to think they are engaged in some kind of rational process of giving each other arguments). So, given that you *already have to somehow account for reasoning about fictions or ceterus paribus clauses*, why not appeal to that here? For both of these areas look like they could provide exactly the right kind of weakening we intuitively want. Isn’t it true that returning weapons to madmen is, ceterus paribus, unjust? Or consider the fiction ‘I lend weapons to someone. But then they go mad. A few

days later, they ask for their weapons back, and I give them back'. Plausibly, it's true in that fiction, that I have paid my debt, but done something unjust.

All in all, one might wonder what's going on here. Why does Williamson appeal to counterfactuals if they don't seem to do any work? And why doesn't he even consider appeals to *ceteris paribus* clauses, and fictions? I suspect that both are motivated by the laudable desire to find some kind of mechanism at work in philosophical reasoning about thought experiments, and then to give an account of why this mechanism should be remotely reliable.

Counterfactuals are especially good for these purposes, because (as Williamson stresses in the previous chapter) our methods of reasoning about counterfactuals can be systematically corrected by cases where the antecedent of a counterfactual actually comes about, so that we can simply observe whether the consequent holds as well. Appeals to truth in a fiction, or *ceteris paribus* clauses are bad because they a) there isn't any obvious similar way that these judgments could be corrected by experience and b) the task of defining truth in fiction or truth *ceteris paribus* in other terms is infamously intractable - so there's little hope of reducing these judgements to some logical product of judgements of some other kind, whose epistemology might prove more tractable.

However, it's plausible that we can achieve this desideratum, without accepting the bizarre conclusion that when moral philosophers say doing x would be wrong they are acting with (generally unjustified) confidence that no one will ever do x in a situation where other factors make doing x permissible. I won't try to argue for my own preferred account of the epistemology of the *a priori* here. But we can note that Davidsonian ideas about charitable interpretation, if correct, apply just as much to talk about *ceteris paribus* clauses, and fictions just as much as they do to literal talk. Also, scientists clearly do revise *ceteris paribus* laws that lead us to form false expectations about future experience. So

these laws are ‘disciplined’ by experience in a sense, even if there are no formal, syntactically specifiable, criteria for when the ceterus are paribus, so that failure to observe relevant behavior means the ceterus paribus law in question must be revised.

In this paper, I have argued that Williamson’s proposed counterfactual account of arguments by thought experiment, doesn’t solve the problem which he introduces it to solve. Whether you take the straightforward indicative account he starts with, or his counterfactual variant, it’s still the case that intuitively acceptable arguments-by-thought-experiment (e.g. the Socrates’ argument that justice isn’t paying your debts), turn out to have implausible empirical commitments (there will never be a real-life situation in which it is just to return weapons to a madman). Nor does he provide any reason to prefer the counterfactual account to alternatives (like fictional or ceterus paribus accounts) which would solve this problem.