Philos Stud (2009) 145:435–443
DOI 10.1007/s11098-009-9403-9

Knowing the intuition and knowing the counterfactual

Jonathan Ichikawa

Published online: 11 April 2009

Abstract I criticize Timothy Williamson's characterization of thought experiments on which the central judgments are judgments of contingent counterfactuals. The fragility of these counterfactuals makes them too easily false, and too difficult to know.

Keywords Thought experiments · Intuition · Timothy Williamson

1 Thought-experiment intuitions as counterfactual judgments

In our (2009), Benjamin Jarvis and I offered a criticism of Williamson's (2005) approach to thought experiments,[1] and offered an alternative account. Williamson's approach to thought experiments then was the same as the one that is developed in his book: the central premise in a thought-experiment based argument is a counterfactual judgment about what *would* be the case if the description provided in the text of the thought-experiment *were* true.

As Williamson (2005, p. 5; 2007, pp. 183–186) puts it, with regard to the Gettier argument:

(2) Possibly, someone could stand to some proposition in the way described by the Gettier text.

[1] We submitted our paper in 2006; we made limited revisions in 2007, reflecting a draft version of *The Philosophy of Philosophy* that was informally available at that time. But we did not then have the chance thoroughly to engage with this more authoritative version of Williamson's project. I am therefore grateful for this opportunity to discuss that project further, in light of both subsequent study of the book, and subsequent conversations with Williamson.

J. Ichikawa (✉)
Arché Philosophical Research Centre, The University of St. Andrews, St. Andrews, Fife, Scotland, UK
e-mail: ichikawa@gmail.com

(3*) If someone were to stand to a proposition in the way described by the Gettier text, then she would have justified true belief in that proposition, but would not know it.

(4) Therefore, it is possible to have justified true belief in a proposition without knowing it.[2]

In a way, Williamson's approach is both unsettling and liberating. It is unsettling, because it renders the central premise, (3*), as a contingently true counterfactual. It is contingent because the description of the story itself does not typically entail justified true belief or non-knowledge; one may fill it out in a variety of ways, only some—including, one hopes, the intended one—result in non-knowledge justified true belief.[3] This represents a radical departure from prior philosophical theorizing about the nature of thought experiments. For example, there is no place, in our judgment that (3*) is true, for anything like George Bealer's (1997) 'feeling of necessity' or Ernest Sosa's (1997) 'abstract contents'.

Radical as it is, however, this approach is also liberating. Indeed, the surprising commonplaceness of the counterfactual judgment, and the vindication of the methods that philosophers use, are two sides of the same coin: *because* on the present approach, philosophers need not engage in any special, distinctively philosophical method in order to run the Gettier argument, worries about the legitimacy of philosophical investigation are made less pressing. If I think that our knowledge of the Gettier premise is the product of a special faculty of rational intuition, then I face the challenge of explaining the source and the reliability of that faculty; if I think it is a commonplace instance of a very general capacity to evaluate counterfactuals, then I need confront no special challenge, if all parties are committed to our possession of this general capacity.

2 When the counterfactual fails

There is certainly something desirable about that liberation. But the radical implication goes too far. Williamson's version of the Gettier premise is *too* contingent. Jarvis and I argued that on Williamson's view, it is too difficult to know the Gettier premise; relatedly, it is also too easy for it to be false.[4]

The counterfactual (3*) is contingent because it is possible that in the nearest worlds where the Gettier text is true, the subject does know. Only if the actual world is not so positioned in modal space is (3*) true. If one happens to be positioned awkwardly in modal space, then one, in running the Gettier argument, relies on a falsehood in premise (3*). Here is an example.

[2] I paraphrase Williamson's approach, which is given in logical notation, into English. A portion of Williamson's chapter, along with his Appendix 2, is devoted to a discussion of the logical formulation of the counterfactual (3*); I circumvent that discussion, as I believe it peripheral to the methodological questions that concern us here.

[3] See Williamson (2007), p. 185, and Ichikawa and Jarvis (2009), p. 224.

[4] Ichikawa and Jarvis (2009), p. 226.

Suppose that one's thought-experiment is given thus:

> At 8:28, somebody looked at a clock to see what time it was. The clock was broken; it had stopped exactly twenty-four hours previously. The subject believed, on the basis of the clock's reading, that it was 8:28.

This should be recognizable as a prototypical Gettier description.

Now consider a world in which that description is true, but where the subject knew in advance that the clock had stopped exactly 24 hours previously. In that world, the Gettier text is true but misleading: its subject knows. So (3*) is false in that world.[5] Someone running the Gettier argument in that world, then, relies on a falsehood, even if he is innocently ignorant of the person who happens to render his counterfactual false. Relatedly, in running the Gettier argument, one commits oneself to being in a world *not* positioned in a way that falsifies (3*). I take these implications to be implausible. (Does one fail to know the Gettier conclusion by virtue of there being someone in his world who satisfies the text in the wrong way?) This was the central argument against Williamson's approach in our (2009).[6]

This result should not be surprising. The JTB account of knowledge is a necessity claim; any possible instance of justified true belief without knowledge refutes it. Williamson's counterfactual is unnecessarily strong: it demands that we find possible instances suitably close in modal space to make true the relevant counterfactuals; this troublesome requirement is unmotivated.[7]

Chapter 6 of *The Philosophy of Philosophy* includes a discussion of the kind of worry that Jarvis and I were pressing.[8] Williamson's response is twofold. He begins by suggesting that quantifier domain restriction can play a helpful role. He writes:

> We might alleviate the problem by understanding the quantifiers in the formalization (3*) … as restricted by the conversational context. For example, it might sometimes exclude instances of the Gettier case on Alpha Centauri.[9]

As Williamson goes on directly to note, this move cannot rescue the Gettier counterfactuals from all scenarios of the sort I am discussing, for "even the contextually relevant domain may happen to betray our expectations." He therefore devotes a greater emphasis to his second strategy, which is to attempt to render plausible the idea that, in cases where the actual world betrays our expectations in a ways that renders (3*) false, we *are* wrong in relying on the Gettier intuition.

I have two concerns about this second strategy: first, that it seems to misdiagnose what happens when actuality betrays our expectations, and second, that it could not,

[5] This is so on a standard Lewis–Stalnaker semantics for counterfactual conditionals, and on any account in which A & $\sim$C entails the falsity of A$\Box\rightarrow$C.

[6] Anna-Sara Malmgren, in a yet-unpublished manuscript, gives an (independently developed) objection to Williamson along similar lines.

[7] Thanks to Crispin Wright here.

[8] pp. 200–204.

[9] p. 200. The quantifiers Williamson here mentions range over subjects and propositions, not over worlds; Williamson's attempt to characterize the counterfactual is:

(3*) $\exists x \exists p\, GC(x,p) \Box\!\!\rightarrow \forall x \forall p [GC(x,p) \supset (JTB(x,p) \& \sim K(x,p))]$.

even if its diagnosis is correct, provide a response to one of the worries presented above. After presenting these concerns, I will consider whether the first strategy—the one relying on quantifier domain restriction—can do the remaining needed work for Williamson.

3 Admitting we are wrong

Williamson has recognized that (*modulo* domain restriction considerations), his view has the counterintuitive consequence that people who coincidentally falsify the Gettier counterfactual undermine the Gettier argument. He attempts to ameliorate this consequence by pointing to a general human tendency to refuse to admit error after relying on premises that turn out to be false in inessential ways:

> Suppose that someone says "Every man in the room is wearing a tie"; I look around, see a man not wearing a tie, misidentify him as Dave (who is in fact wearing a tie), and say "Dave is not." When it is pointed out to me that Dave is wearing a tie, I deceive myself if I insist that my original reply was correct because the man whom I had in mind was not wearing a tie.[10]

Williamson's judgment about this story strikes me as correct. I am not at all convinced, however, that it is analogous to the case of a person who runs through a Gettier argument based on a text, situated in a world in which the text happens to be true in a way that is not a counterexample to K = JTB. If my student considers the clock story above, and comes, in the natural way, to believe the JTB theory of knowledge to be false, it seems to me that she commits herself to no contingent claims about whether there are any real people looking at broken clocks that read 8:28 at 8:28, or about what else would have been true of them, if there had been.

When Williamson's character says "Dave is not," he is committing himself to the world being such that Dave is not wearing a tie. If he insists that his utterance of "Dave is not" really meant that that guy over there was tieless, he does deceive himself; what he said, and relied on, was false, even though there was a truth 'in the neighborhood' that would have done just as well.

By contrast, when my student says (whether out loud, or to herself), "the subject has justified true belief but does not know," i.e., when she expresses the Gettier intuition, she is not, it seems to me, committing herself to the world being such that the nearest worlds in which someone, at 8:28, looks at a broken clock that reads 8:28, are worlds in which he has justified true belief without knowledge. If it is pointed out to her that there happens to be someone with such a broken clock who does know, she is right to react with indifference: "that case is importantly different from the one I was considering."

What Williamson's character expresses by "Dave is not" is the falsehood that Dave is not wearing a tie, even if the character protests that that's not what he meant. But what my student expresses by "the subject does not know" is *not* the falsehood that (3*); it is some other truth. She is right that this rebuttal of (3*) is no

[10] p. 201.

rebuttal to what she said. So (3*) is a poor characterization of the content of the Gettier intuition.

Williamson suggests that the appropriate response to the discovery that one is in a world that is deviant with respect to the Gettier counterfactual is to change the Gettier story: instead of engaging with the possibility that someone looks at a broken clock in the way described above, consider the possibility that someone looks at a broken clock in the way described above, *and* did not previously know that the clock had stopped 24 h previously. We should admit, he says, that the argument we had relied on was unsound, and replace it instead with this alternate, sound argument.

Perhaps Williamson's suggestion enjoys some intuitive support with this verdict. It is fairly natural to respond to this sort of challenge with a further spelling-out of the case to be considered. However, we should not be too quick to judge, on the basis of such data, that the pattern Williamson is suggesting is the correct one. For it is not only in deviant, counterfactual-falsifying worlds that this pattern obtains. It also obtains when the mere *possibility* of such a confounding factor is mentioned, even when it is not, and is not being suggested to be, what would be the case if the description held. Suppose I run the Gettier argument based on the clock story above, and somebody responds by pointing out that the subject *could* have known in advance that the clock had stopped exactly 24 h previously. I may well react to my unwelcome interlocutor by simply modifying the case, changing the argument to one about someone who reads a clock in the way described, and who also did not know in advance that the clock was broken. That is to say, I will react to him in just the same way that I react to the person who points out that I am in a deviant world, as in the previous cases considered.

In these two scenarios just described, I am challenged in two different ways. One of the challenges falsifies the counterfactual (3*); the other does not.[11] But the intuitive response to each is the same: to modify the story and stipulate the objection away. Therefore, that intuitive response does not speak in favor of the falsity of the Gettier premise. Sometimes, we respond in that way even when, by Williamson's lights, the truth of the premise is unthreatened.

4 Knowing (3*)

Setting aside the argument of the previous section, there is another problem with Williamson's account that stems from the failure of the counterfactual in uncooperative worlds. If we are convinced by Williamson's argument that we

[11] This is very straightforward on anything in the ballpark of Lewis's standard (invariantist) semantics for counterfactuals. Things are a bit more complicated on a contextualist semantics, but the ultimate result is the same. Consider some possibility D (*deviant*), where D together with G (*Gettier*) result in the negation of C (*counterexample*). On a contexutalist view, conversational salience of D might make an utterance of 'G$\Box\rightarrow$ C' false, even though, had D not been salient, such an utterance would have been true. It is still not the case, however, that the subsequent discussion of D challenges the truth of *what was said* by the earlier utterance of 'G$\Box\rightarrow$ C'. What I said was true, though if I used the same words again now I had say something false.

should reject our central premise as false when the world is uncooperative, then perhaps we can explain the way to engage with the discovery that the counterfactual is false in these deviant worlds. However, an epistemic problem remains: even in normal worlds, where the counterfactual is true as intended, how do we *know* that the counterfactual is true? I believe that Williamson's account renders it much too difficult to know the Gettier intuition.

It is not enough to say that we can know (3*) through a general capacity to evaluate counterfactuals. One may admit the general capacity to evaluate counterfactuals, while remaining skeptical about one's ability to know a particular counterfactual. (I think that Tim has a general perceptual ability to identify the colors of objects, but also that he has no way to know what color shirt I am wearing right now.)

The worry is that in too many cases, it is not plausible that we know the relevant counterfactuals, because it is not plausible that we know whether we are situated deviantly in modal space. The world is a big place, and we should not be at all surprised if it turns out to be true that someone has been in a broken clock situation in the way described above. And I see few prospects for learning whether, in all cases in which it is happened (if it is happened), or in the nearest cases in which it's happened (if it has not), it happens in a deviant way that prevents it from being a counterexample to K = JTB.

Obviously, the more specific the story, the less pressing the worry becomes. If we are using Crispin Wright's (1983) Gettier case about John McEnroe at Wimbledon, we may, if we know enough about the history of tennis, be very confident that the description hasn't been satisfied in a deviant way. My clock story is a more difficult one; a more abstract clock story, in which the particular time isn't mentioned, is more difficult still. But all of these stories function equally well for establishing knowledge of the conclusion of the Gettier argument. The McEnroe counterfactual is not very difficult to know; the no-time-mentioned clock counterfactual is, even if it happens to be true. (I have no idea whether it is true, or even how to guess how likely it is to be true.)

It would appear to be an implication of the view, then, that the McEnroe version of the argument is a much better one than the clock version, since most people know the former counterfactual but not the latter. But this is not the case; both thought experiments establish knowledge of the Gettier conclusion equally well. This suggests that proper execution of the Gettier argument does not require the knowledge of the counterfactuals in question, and that those counterfactuals, therefore, are not identical to the Gettier intuition.

5 Domain Restriction

In response to these worries, perhaps Williamson will want to rely more heavily on the domain restriction move.[12] Can it do the remaining work?

[12] At the Arché workshop on which this symposium is based, Williamson responded to my presentation of such epistemic worries by invoking domain restriction.

Sometimes, universals are true, even though there are would-be counterexamples outside their restricted domains. I say that I am grouchy because all my friends are tired of my new joke. You cannot refute me by reminding me that I have a friend in Cambodia who has not yet heard my new joke; I was talking about my friends at this dinner party. If you do so remind me, I need not take back my original claim. Nor must I modify it to "all my friends *at this party* are tired of my joke," although I might say that amended sentence as a way to clarify my earlier claim. (So domain restriction cases provide another class of cases like the ones alluded to above, where the most expedient response to a challenge may be to change what we say, even when what we said the first time was perfectly true.)

As discussed above, Williamson argues, plausibly, that philosophers are prone to over-apply this sort of move, wrongly insisting to have been right in the first place even in cases where the original claim was literally false. But if we are to take the phenomenon of semantic quantifier domain restriction seriously, then this, at least, should be considered a case in which the original utterance about "all" my friends was true.

I wrote a very powerful poem, but burned it before anyone heard it. If someone had heard it, he would have found it moving. Some people are excluded from the domain of the quantifiers characterizing that counterfactual. I did not mean to be referencing my sociopathic roommate, for instance. He never finds anything moving. Maybe, the way I happen to be situated in modal space, he is the only person who almost heard it; my counterfactual could still be true if its domain excluded him. (Maybe he *did* hear it and the 'anyone' quantifier in the first sentence also excluded him.)

For my own part, I admit to uncertain intuitions about whether my sociopathic roommate falsifies the counterfactual. But let us suppose, with Williamson, that this sort of thing can happen in some cases. It will not, I think, be enough to resolve the worries that I have been developing.

The quantifiers Williamson suggests restricting range over people, not possible worlds. But the deviant possibilities I have been discussing, which threaten to undermine (3*), depend on weird situations, not on weird people. We can exclude certain classes of people from the domain of the counterfactual involving clocks: people on Alpha Centauri, infants, insane people, etc. None of this solves the problem, since even very ordinary people can, if the world happens to be the wrong way, falsify the counterfactual. In the cases we have been imagining, there has been nothing particularly unusual about the subjects at all. They are not the sorts of people we meant to be leaving out of our quantifier domains.

Of course, there is one way to restrict the quantifiers in a way that avoids the problem: we could let our quantifiers range only over the people who are not situated in the world so as to have justified true belief without knowledge, in the nearest possibility where they satisfy the Gettier text. Anyone who would fail to be a counterexample to K = JTB, were the description true, is excluded from the domain. This feels a bit like cheating. One may have doubts about whether the metasemantic rules underlying the mechanics of quantifier domains render such a restriction plausible. But this move is unavailable to Williamson for another reason: as discussed above, he does think that there are at least some cases where the world

conspires to falsify the counterfactual; if the domain were restricted in the robust way described in this paragraph, the world could never conspire against it. If we restrict the domain to people who would, if they were in a Gettier case, have justified true belief that is not knowledge, then is clear that *anyone* in a Gettier case would have justified true belief that is not knowledge. This, it should be clear, is not close to Williamson's view.[13]

6 Reasoning with Richer scenarios

This sort of super-restricted quantifier move has some elements in common with the proposal that Jarvis and I offered to replace Williamson's. In both cases, the sorts of deviant possibilities that have been concerning us are not, even if they happen to be nearby or actual, genuine counterexamples to the content of the Gettier intuition. We suggested that in running the Gettier argument, one considers the Gettier text, then enriches it, considering a more-determinate scenario, *g*—one that entails justified true belief without knowledge—on which he can then run the Gettier argument in a straightforward way:

(5) $\Diamond g$
(6) $\Box(g \supset \exists x \exists p\,[\mathrm{JTB}(x,p) \& \sim \mathrm{K}(x,p)])$[14]
(4) $\Diamond \exists x\ \exists p\,[\mathrm{JTB}(x,p) \& \sim \mathrm{K}(x,p)]$

One objection that Williamson has offered to this sort of approach is that it denies us of a public argument.[15] Everyone engages with a Gettier case via her own private argument, which depends on how she happens to have filled out the scenario in her own mind.

But our view is not that subjects should fill out thought experiments in whatever way they like; we have particular conventions, grounded in our practices with fictions, that govern how to move from a weaker description to a stronger scenario in the intended way. So there is no problem with discussing thought-experiment scenarios as publically available. (Just as there is a publically available *Matilda* story that is not entailed by the sentences used to generate it.)

It is true that, on our view, there will be possible cases of miscommunication, where the error lies in divergent private fillings-out of the scenario. Perhaps there is

[13] One might go on to worry, along similar lines, whether the universal claim (3) that Williamson rejects (p. 185) could be vindicated by this sort of super-robust domain restriction. However, although it easily proven that, with such a restriction, all Gettier subjects are counterexamples to K = JTB, the necessity claim of:

(3) $\Box \forall x \forall p[\mathrm{GC}(x,p) \supset (\mathrm{JTB}(x,p) \& \sim \mathrm{K}(x,p))]$

is still false. Although the subjects in the domain are, by hypothesis, not actual counterexamples to (3), there are worlds where they are deviantly situated such that they are counterexamples. (Compare: where the domain of the quantifier consists in all faculty members, 'necessarily, everyone is a faculty member' is falsified by the possibility that Professor Doddery could have retired last year.)

[14] That this necessity claim is true reflects an important difference between our approach and the super-restricted quantifier one discussed above. See previous footnote.

[15] Williamson offered this objection at the Arché workshop in September 2008.

something less than ideal about this state of affairs, but, in my view, this added moving part is just a part of philosophy.[16] Sometimes we have to do extra work to ensure that our thought experiments are being thought of the way we intend.[17]

Acknowledgment I am grateful for the privilege of engaging with Timothy Williamson's fascinating book, and in such distinguished company.

References

Bealer, G. (1997). Intuition and the autonomy of philosophy. In M. DePaul & W. Ramsey (Eds.), *Rethinking intuition*. Lanham, MD: Rowman & Littlefield.

Ichikawa, J., & Jarvis, B. (2009). Thought-experiment intuitions and truth in fiction. *Philosophical Studies, 142*(2), 221–246.

Sosa, E. (1997). Minimal intuition. In M. DePaul & W. Ramsey (Eds.), *Rethinking intuition*. Lanham, MD: Rowman & Littlefield.

Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies, 132*(1), 99–107.

Williamson, T. (2005). Armchair philosophy, metaphysical modality and counterfactual thinking. *Proceedings of the Aristotelian Society, 105*(1), 1–23.

Williamson, T. (2007). *The philosophy of philosophy*. Malden, MA: Blackwell.

Wright, C. (1983). *Frege's conception of numbers as objects*. Aberdeen, Scotland: Aberdeen University Press.

[16] Sosa (2007) suggests that some apparent disagreement about philosophical thought experiments might be explained in this way.

[17] I presented a version of some of this material at an Arché workshop at the University of St. Andrews, devoted to Williamson's excellent book. I am grateful to the participants there, and to the AHRC, which funded the event. Thanks also for helpful comments on an earlier version of this paper to Anna-Sara Malmgren and Crispin Wright. I am especially indebted to Benjamin Jarvis, who co-authored the (2009) paper with me, and to Timothy Williamson.