Topoi (2019) 38:801–810
DOI 10.1007/s11245-017-9491-7

# Thought Experiments, Formalization, and Disagreement

**Sören Häggqvist**[1]

Published online: 5 July 2017


**Abstract** In the last decade, philosophers have offered a number of proposals concerning the logical form of hypothetical cases, or thought experiments, as these are used for purposes of testing philosophical claims. In this paper, I discuss what the desiderata for a formal proposal are. Employing a comparison with general philosophy of science, I suggest that one important desideratum is to highlight recurrent patterns of disagreement surrounding cases. I advocate a proposal in propositional modal logic which, I argue, better meets this desideratum than competing proposals. I also sketch how this proposal may be extended into more fine grained analyses, employing counterfactual conditionals yet avoiding certain problems due to so-called "deviant realizations".

**Keywords** Thought experiments · Intuitive judgements · Formalization · Disagreement · Williamson

## 1 Introduction: The Recent Debate

In the last decade, philosophers have offered competing proposals concerning the logical form of hypothetical cases, or thought experiments, as these are used for purposes of testing philosophical claims (Williamson 2005, 2007; Ichikawa and Jarvis 2009, 2013; Malmgren 2011). Debate about these proposals has partly revolved around their putative epistemological consequences: whether they allow intuitive judgements about cases to be a priori, and whether they render such judgements sufficiently reliable, or sufficiently knowable. Some debate has also concerned the proposals' semantic plausibility—that is, whether they capture what people actually judge when confronted with cases.

Both semantic and epistemological criticisms sometimes focus on the so-called "deviance objection". This objection holds that putative renderings of what a thinker judges in response to a case risk becoming false due to unintended, intuitively irrelevant *realizations* of the judgement (as construed by the proposal under criticism). To see how deviance objections work, consider the following Gettier scenario (used by Malmgren 2011, p. 272).

(S) Suppose that Smith believes that Jones owns a Ford, on the basis of seeing Jones drive a Ford to work and remembering that Jones always drove a Ford in the past. From this, Smith infers that someone in his office owns a Ford. Suppose furthermore that someone in Smith's office does own a Ford—but it is not Jones, it is Brown. (Jones's Ford was stolen and Jones now drives a rented Ford.)

(S) naturally invites the judgement that Smith doesn't know that someone in his office owns a Ford. Given that (S) is possible, this entails that justified true belief is not knowledge in all possible worlds.

Williamson's proposal renders this as follows (Williamson 2007, ch. 6):

($i_{GW}$) $\Diamond\exists x\exists p\ (GCx,p)$

($ii_{GW}$) $\exists x\exists pGC(x,p)\ \Box\!\!\rightarrow\ \forall x\forall p(GC(x,p) \supset (JTB(x,p) \ \&\ \neg K(x,p)))$

$\therefore$ $\neg\Box\forall x\forall p(K(x,p) \leftrightarrow JTB(x,p))$,

✉ Sören Häggqvist
soren.haggqvist@philosophy.su.se

[1] Department of Philosophy, Stockholm University, 106 91 Stockholm, Sweden

where the modal operators all indicate metaphysical modality, the JTB thesis is expressed as the necessary equivalence of justified true belief and knowledge, the variable "x" ranges over people and "p" over propositions, and the predicate "GC" is satisfied by persons and propositions related just as in (S). The first premise asserts the possibility of the scenario. The second premise is the Gettier judgement, here a counterfactual roughly to the effect that if someone were related to a proposition *p* as Smith is in (S), then she would have a justified true belief that *p* but would not know that *p*.[1] Together the premises entail the falsity of the JTB thesis as construed.

One deviance objection against Williamson observes that ($ii_{GW}$) is false if *someone* happens to have information undercutting her justification of a proposition to which she is nonetheless related in exactly the same way as Smith is in (S) to the proposition that someone in his office owns a Ford. Suppose there is such a person, who in addition to satisfying the descriptions explicitly stated in (S) also has "good reasons to believe that [s]he is prone to hallucinate people driving Fords to work and prone to misremember what cars people drove in the past" (Malmgren 2011, p. 279). If such a person exists somewhere—even unbeknownst to the thought experimenter or contemplator—this renders ($ii_{GW}$) false. Similarly, someone may be in the relation dictated by (S) but consistently with this have some alternative route to knowledge of the proposition to which she is thus related.

The semantic version of this deviance objection is that the thought experimenter's judgement doesn't pronounce on such people, and that ($ii_{GW}$) hence misrepresents the judgement. The epistemic version objects that in order to know the crucial judgement about (S), one needn't know that no such deviant realizations of (S) obtain, whereas this is required by ($ii_{GW}$). Hence ($ii_{GW}$) incurs unwelcome epistemic risks.

In Sect. 5 I shall suggest that one source of such deviance simply is Williamson's choice of representing apparent singular terms with bound variables in his formal proposal. This is not mandated by rendering judgements about cases as counterfactuals (Sect. 6). But before this point can be put in its dialectically proper place, the question of what we should ask from a formal proposal about thought experiments must be considered. In discussing the aims of formalization (Sect. 2), I suggest that we take a pragmatic approach—more pragmatic, that is, than that of the proposals under discussion here. I will argue that an important desideratum is a proposal's ability to display a common feature of a significant class of cases, and that one such feature is that there are certain recurring ways of disagreeing with the intended conclusion of philosophical thought experiments (Sect. 3). I complain that the proposals from Williamson, Ichikawa, Jarvis, and Malmgren fail to meet this desideratum (Sect. 4). Besides returning to the issue of deviance, the concluding sections sketch why I believe that a proposal couched in propositional modal logic I have defended elsewhere (Häggqvist 1996, 2009) meets the desideratum I stress. In Sect. 6, I briefly suggest how this proposal may be elaborated into more finegrained QML renderings of thought experiments.

## 2 Why Formalize?

General philosophy of science may offer some leads on this question. Formalization of the "logic of hypothesis testing", as carried out by e.g. Hempel and Popper, wasn't in the business of vindicating experimentation. Flawed and successful falsifications alike may be represented in similar fashion. What formalization offered was a useful general format for better understanding the reasoning involved in scientific inference, and in particular certain common ways in which a putative falsification might be challenged and contested. However, formalization is quite impotent when it comes to warranting particular experiments (or tests generally). Such warrant will depend on particulars of different experiments that usually will not be displayed in the general formal format. For instance, whether the auxiliary hypotheses suffice for inferring the pertinent predictions, whether they are more likely than the falsity of the theory, whether the initial conditions were correctly established and whether the observations were correct or artifacts; and so on.

Yet such formalization is not epistemically useless. It provides a taxonomy for seeing broad commonalities among, and thus having a better understanding of, experiments in general.[2] One area of such commonalities concerns the broad, recurrent ways in which an experiment may be contested. In this way, formalization of scientific inference offered a coarse but useful taxonomy of scientific controversy (among other things).

One aim of formalizing thought experiments—in the sense operative here and in the three proposals mentioned above, namely, as modal counterexamples to general theories—might reasonably seek to do for them something analogous to what formal philosophy of science did for ordinary hypothesis testing.

[1] See Williamson (2007, pp. 195–199) for discussion of whether ($ii_{GW}$) sufficiently captures the anaphoric reference of this English sentence.

[2] As well as providing a better understanding of salient differences between confirmation and falsification, of course.

In order to achieve such an aim, generality—i.e. wide applicability to many instances—is important. Of course, generality comes at a price: a formalization applies to lots of instances by riding roughshod over their peculiar details. But that's fine as long as it is understood that the aim is not to identify an objectively established unique form, let alone any psychologically real mental contents of individual researchers, but rather just to come up with a reconstruction useful for broad understanding. Consider Quine on logical form: "there need be no question of the uniquely right analysis" (1960, p. 160). On his view, logical form is to some extent imposed for certain purposes rather than found. Of course, it cannot be imposed gratuitously: there has to be some fit with what is formalized. But the fit may be somewhat procrustean. This attitude to logical form seems to have been widely accepted among twentieth century general philosophers of science. And it seems sensible for purposes of formalizing thought experiments, too.

By contrast, recent formal proposals for thought experiments are coupled with an epistemological agenda. They seek to vindicate thought experiments, and often (Williamson being an exception) this is seen as part of a wider aim of vindicating rationalism. Malmgren goes even further, and states that an aim of formalization is to capture the psychologically real contents of (judgements about) cases:

> In this paper, I defend rationalism against a recent objection, due to Williamson, that threatens to undermine the prima facie case [for rationalism]. In so doing, I discuss the formal structure of thought experiments in more detail; in particular, how to analyse intuitive judgements—what their 'real' content is (Malmgren 2011, p. 266)[3]

She explicity rejects the aim of providing a mere rational reconstruction (Malmgren 2011, p. 283). Questions of psychological plausibility similarly inform the proposal due to Ichikawa and Jarvis (2009). By contrast, I suggest that we take a more pragmatic approach, namely one in which there needn't be a single right form and in which questions of psychological realism do not enter at all.[4]

Another issue concerns logical grain. The proposals put forward by Williamson, Malmgren, and Ichikawa & Jarvis are all couched in quantified modal logic. Notably, while explicitly aspiring to generality, they are also all expounded with detailed reference to, and only to, Gettier cases. In the next section, I shall explain why I think an excessive focus on Gettier cases can be misleading if one seeks a broader understanding of the workings of thought experiments in philosophy. But it is interesting to consider whether someone, like myself, who takes a more pragmatic approach to the form of thought experiments should aim to couch a proposal in QML as well. Again, I believe that the parallel with formal regimentation in the philosophy of science may be instructive. In the latter, quantified formalization may be required to see how, in general, a simple universally quantified statement can entail an observation conditional. But for other purposes—e.g. seeing how an overall argument is valid in a falsification, and that confirmation is not analogous—propositional logic suffices. I think the same holds for thought experiments. Some general facets may be displayed by a coarse-grained model in propositional modal logic. Accordingly, I will argue, for purposes of discussing and understanding general features of thought experiments, a PML proposal is often enough.[5]

## 3 Patterns of Disagreement and Gettier Myopia

Recent proposals in this area are preoccupied with Gettier cases (Gettier 1963). Why? As noted, these proposals all seek to vindicate thought experimentation, and at least historically, Gettier cases have been dialectically successful. Williamson notes that Gettier's "refutation of the justified true belief analysis was accepted almost overnight by the community of analytic epistemologists" (2007, p. 180). He offers as his "background working hypothesis … that [Gettier's] thought experiments are paradigmatic, in the sense that if any thought experiments can succeed in philosophy, his do: thus to determine whether Gettier's thought experiments succeed is in effect to determine whether there can be successful thought experiments in philosophy" (Williamson 2007, pp. 179–180). Later authors have viewed Gettier cases as paradigmatic in a stronger (and more standard) sense.[6] Ichikawa and Jarvis thus simply note that

[3] In other places in her paper, the scare quotes around "real" are dropped.

[4] Malmgren's demand for psychologically real contents seems to me elusive. Psychologically real contents of an individual thinker's judgement about a given case might, for all we know, differ even intrapersonally on different encounters with the same case (even if their verbal expression is the same). Such contents would also seem difficult to establish without the help of some empirical (psychological) investigation. Finally, the aim for the specific, real *content* of a given (token?) judgement seems to be in tension with the aim of general application to several different cases.

[5] This is not to deny that the finer grain of QML may be desirable if we want to understand, and represent, the finer details of debates surrounding a particular thought experiment. Again, I am with Quine: "A maxim of *shallow analysis* prevails: *expose no more logical structure than seems useful* for … the inquiry at hand" (Quine 1960, p. 160; italics in original), provided (the natural implicature) that we not expose less structure than is needed.

[6] Besides Malmgren (2011) and Ichikawa and Jarvis (2009), texts treating Gettier cases as a in some ways paradigmatic of philosophical thought experiments include e.g. Chudnoff (2011), Ichikawa (2009), and perhaps Nagel (2013). But it isn't plausible to claim that they were so regarded in earlier literature. In one agenda-setting col-

they "follow Williamson in using the Gettier intuition as a paradigm" without qualifying "paradigm" (2009, p. 223). This is of course unobjectionable in authors motivating their own proposals by criticizing Williamson's treatment of Gettier cases. But if the aim is to give a proposal covering lots of different cases, as these proposals all aspire to do, there is some reason to regard Gettier cases as unrepresentative, or at least to consider a range of dialectically less successful instances. It is then striking that many philosophical thought experiments are *contested* to a much greater extent than Gettier cases.

Contestation doesn't reduce to "variation in intuitions", although that is certainly one source for contesting cases. Rather, the point is that there are various logically coherent ways of disagreeing with the claim that a certain thought experiment shows a certain target theory false—or, for brevity, of disagreeing with a thought experiment. Often, these ways of disagreeing are given professional expression (here, Gettier cases are a notable exception). And often, they are both coherent and not entirely implausible. This trite observation has its exact parallel for ordinary empirical experiments. It doesn't in the least impugn the feasibility of successful thought experimentation.

But an interesting fact about such disagreements, which we might want a formal proposal to display, is that different ways of disagreeing with thought experiments exhibit interesting similarities *across* cases, and are thus naturally grouped into a few broad patterns. This is just the sort of thing formal representation is good for. Hence, if our interest is to understand thought experiments more generally—as opposed to understanding how they succeed epistemically whenever they do—it seems desirable that we achieve some grasp of their structure that is useful for understanding and discussing various ways in which thinkers may disagree with their intended conclusions. Gettier cases are not apt to help us see the range of available patterns here, I surmise.

I think that there are three main ways of disagreeing with a thought experiment. The first is plain: disagreeing with the presented claim about the case—in Gettier cases, this would be tantamount to judging that the protagonist of a case does know. Disagreement of this type—which we may call "outcome disagreement"—is exceedingly rare among professional philosophers in Gettier cases.[7] A second type of disagreement concerns the possibility of the scenario presented in a thought experiment's vignette. Again, few would deny that Gettier scenarios are possible.[8] But to see that both possibility disagreements and outcome disagreements are prevalent, and often tenable, in philosophy, one just needs to broaden one's gaze a little.

Consider Searle's (1980) Chinese Room, offered as a counterexample against what Searle calls "strong artificial intelligence". One way of disagreeing with Searle's thought experiment is to hold that the person in the room would understand Chinese, or that the system comprising what is in the room would. But another way is to deny that the scenario is possible—e.g. because the stipulation about how the room works is incompatible with stipulations (on one of its versions) about its speed (Dennett 1987). Critics of Searle may vacillate between these two responses (in part because different specifications of the case may motivate different responses). But it would be wrong to view these as the same response. They are two different ways of disagreeing with Searle. And they exemplify recurrent ways of disagreeing with thought experiments.

However, besides these two common types of disagreement with the intended conclusion of a thought experiment, there is a third. Consider, as we have done throughout, negative experiments intending to refute some general claim. In certain of these cases, one may agree that the scenario is possible, agree with the intended intuitive judgement about the case, and yet deny that this refutes the general claim that is targeted for refutation. This is, in effect, to deny that the experiment is relevant (so we may dub this type "relevance disagreement") for the theory under testing, although its scenario is logically, perhaps even metaphysically possible.

But how could a scenario be irrelevant to a theory, yet possible? A key to seeing this is to abandon an assumption implicit in most recent theorizing: that theories tested by means of thought experiments are claims of metaphysical necessity. Again, Gettier cases are of little help, since it is quite plausible to construe the JTB theory as a metaphysical necessity claim (as Williamson, Malmgren, and Ichikawa & Jarvis all do). Other theories and claims subject to testing by thought experiments, however, are presumably

Footnote 6 (continued)

lection (Horowitz and Massey 1991), Gettier cases are barely mentioned. Monographies in the following years, such as Sorensen (1992), mention Gettier cases but don't suggest any special status for them; the same holds for two major treatments in German (Kühne 2005; Cohnitz 2006). The very influential Weinberg et al. (2001) uses Gettier cases as probes alongside a number of other epistemological thought experiments, such as Fakebarn cases, Zebra cases, and Conspiracy cases, but these are all placed on an equal footing.

[7] An exception may be Weatherson (2003), although his reasons for judging contrary to orthodoxy are avowedly theory-driven; he admits to a strong inclination of denying knowledge in these cases. Among laymen, there is some startling apparent disagreement—see Weinberg et al. (2001) and, for some plausible potential explanations, Nagel (2013).

[8] Except on (controversial) metaphysical grounds, assuming that the scenario is a fiction while invoking the Kripkean claim that ficta exist in no possible worlds. Cf. Ichikawa and Jarvis (2009, p. 229, fn. 13).

neither intended nor reasonably taken to aspire to metaphysical necessity. There is, for instance, nothing in Thomson's writings on either trolley cases (Thomson 1973, 1985) or on her famous violinist case (Thomson 1971) to suggest that she envisages the claims against which she is arguing as metaphysically necessary. The same holds for Searle and "strong AI". Arguably, it holds even for Putnam's (1975) Twin Earth case.[9]

In general, I submit, thought experiments are widely used in areas of philosophy where the theories under discussion simply aren't intended to be true in all metaphysically possible worlds. Political philosophy and philosophy of mind are two such areas.[10] In these areas, theories may have smaller "modal scope".[11] Just what the modal scope of a theory is is sometimes itself subject to debate. Such debate, which may have direct influence on the evaluation of a thought experiment's significance, needn't be based on fallacy or confusion. Hence relevance disagreement is a move whose in-principle availability we have pragmatic reason to display in a formal proposal.

## 4 Capturing Patterns of Disagreement

If we accept that the patterns of disagreement I've mentioned are worth capturing, it will be instructive to see to what extent recent proposals can do so. To this end, let me briefly summarize the two proposals of Ichikawa and Jarvis (2009) and Malmgren (2011) (Williamson's was exhibited in Sect. 1).

*Pace* Williamson, Ichikawa & Jarvis propose that the Gettier judgement is a strict conditional. But in their view, the antecedent of this conditional is best construed as a fiction or story, for which the explicit vignette is just a starting point, to be "enriched" or filled out in the same way as we usually enrich or fill out fictional narratives. Thus the Gettier fiction, $g$, resulting from enriching a vignette such as (S), can be understood as a proposition, i.e. a set of worlds where this fiction is true. What orthodox contemplators of a Gettier case judge is that all members of this set contain an instance of someone having justified true belief without knowledge. On Ichikawa & Jarvis's proposal, then, the argument is

$(i_{GIJ})$ $\Diamond g$

$(ii_{GIJ})$ $\Box(g \supset \exists x \exists p(JTBx,p \ \& \ \neg Kx,p))$

$\therefore$ $\neg\Box\forall x\forall p(K(x,p) \leftrightarrow JTB(x,p))$,

where, as before, the operators are metaphysical and the variables range over people and propositions, respectively.[12]

Malmgren's proposal is a direct inference from the Gettier judgement, as she construes it, to the same negative conclusion as in the other proposals (Malmgren 2011, p. 281):

$(i_{GM})$ $\Diamond\exists x\exists p\ (GC(x,p)\ \&\ JTB(x,p)\ \&\ \neg K(x,p))$

$\therefore$ $\neg\Box\forall x\forall p(K(x,p) \leftrightarrow JTB(x,p))$.

Now the distinction between outcome disagreements and possibility disagreements is readily captured by accounts, such as Williamson's and Ichikawa's and Jarvis's, that partition the premises for the negative conclusion into separate propositions where one explicitly claims that the scenario is possible. On Malmgren's account, the best way of rendering a denial that the scenario is possible (implausible as this would be in a Gettier case) seems to be "It is impossible that someone stands to $p$ as in the Gettier case (as described)", or, symbolized:[13]

$(PD_M)$ $\neg\Diamond\exists x\exists p\ (GC(x,p))$,

and the best way of rendering an outcome disagreement (again, implausibly in a Gettier case) seems to be

$(OD_M)$ $\neg\Diamond\exists x\exists p\ (GC(x,p)\ \&\ JTB(x,p)\ \&\ \neg K(x,p))$.

On this proposal, a possibility disagreement entails an outcome disagreement. More worryingly, $(OD_M)$ appears too strong: it construes a dissenting verdict about the outcome of the particular case as a sweeping denial of the possible existence, throughout logical space, of any pairs

[9] Putnam actually claims never to have endorsed the notion of metaphysical necessity in anything like Kripke's sense (Putnam 2015). Since he specifies this in the context of defending his conclusions based on the Twin Earth Case, he presumably does not take whatever modality attaches to semantic externalism (or internalism) to be metaphysical.

[10] As is, I believe, ethics, although this is less visible now that much of meta-ethics has come under the sway of analytical metaphysics. But cf. Hare (1984) for an instance of a relevance disagreement, of approxiamtely the sort I intend, with cases directed at utilitarianism.

[11] This point extends to the use of thought experiments in science. For a recent argument that generalizations in special sciences aim at modal *robustness* rather than necessity, see Strevens (2012).

[12] The rendering of the argument is mine, but departs only inessentially from that of the authors. Ichikawa and Jarvis add a caveat: since the protagonist of vignettes (such as GC1) are typically fictional, and since fictional characters are, perhaps, essentially fictional in the sense that they exist at no possible world (as Kripke thinks), a technical solution may be needed: introduce vignette characters' "names" via (stipulative) descriptions. (Ichikawa and Jarvis 2009, fn. 13). As they note, Williamson appears to make a similar move after noting that vignette sentences containing fictional names may not express propositions (2007, p. 184). So does Malmgren. I will return to this issue below.

[13] "OD" and "PD" chosen as mnemonics for "outcome disagreement and "possibility disagreement".

of subject and propositions satisfying the three predicates "GC", "JTB" and "¬K". Even bearing in mind that the "GC" predicate here is very specific (since it is standing for the relation specified in a given, specific vignette), and even granting the leeway for procrustean formalization I stressed above, this seems wrong as an attempt at rendering what the dissenter claims.[14]

The chief shortcoming of the three proposals, however, is that they lack the resources for expressing what I called relevance disagreements. To display these, a proposal must avoid construing the relation between the target claim under testing and the case itself as a matter of logical entailment, on the basis of the presumed metaphysical necessity of the target claim. Clearly, it must also employ at least three premises, each a natural target for denial on each of the three types of disagreement.

Here is a simple propositional model I have offered in earlier work (Häggqvist 2009) which meets these requirements:[15]

(i) $\Diamond C$

(ii) $C \mathrel{\Box\!\!\rightarrow} \neg P$

(iii) $T \supset (C \mathrel{\Box\!\!\rightarrow} P)$

$\therefore \quad \neg T,$

where "T" is the claim under testing, "C" is the scenario whose possibility (i) asserts, and "P" is a claim about what would be the case if C were to hold to which T is held committed (iii), but which on the intended outcome of the case, would be false if C were to hold (ii). The point of letting (iii) be as weak as a material conditional is to allow, in principle, a position on which the conclusion is denied by someone who nevertheless grants the outcome claim (ii) as well as the possibility of the scenario (i); in short, someone who voices a relevance disagreement.

Now one basis for relevance disagreements—and a rationale for wanting a proposal to display them as a logically feasible option—is, as I indicated above, that theories tested by cases may not be, or not be recognized as, metaphysical necessities. If C is possible only at worlds outside the modal scope of T, (i) and (ii) may be accepted without forcing the conclusion that T is false: then (iii) is rejected. Thus (iii) expresses a weak claim to which the proponent of a thought experiment, in the present sense, is plausibly committed. But note that (iii) is not itself supposed to carry any substantive information about what the modal scope of T is, let alone express that scope (for that, rather more would be required).[16]

This model meets the pragmatic desiderata I emphasized in Sect. 2.[17] But it is committed to construing "outcome judgements" about cases as counterfactuals. Letting counterfactuals serve as the main connective, as proposed here and by Williamson, has been the target of much recent criticism, often focussing on the "deviance objection" I mentioned in the beginning. The rest of this paper will delve into some aspects of this criticism.

## 5 Deviance

As we saw in Sect. 1, Williamson's candidate for expressing a Gettier intuition is "deviantly realized" whenever the closest antecedent-world—perhaps the actual world—happens to contain a person who satisfies GC(x, p) but either fails to satisfy JTB(x, p) or does satisfy K(x, p) because of factors compatible with, but not mentioned in, the vignette. As construed by Malmgren, the point is semantic rather than epistemological: the worry is not that the judgement as construed by Williamson's proposal may be false, but that its potential falsity due to intuitively *irrelevant* factors shows that the proposal deviates from what the judger meant (cf. Malmgren 2011, p. 276). Clearly, we also shouldn't construe the judgement as *obviously* false—but this is, as Malmgren notes, a distinct issue. Hence, Williamson's strategy for dealing with the possibility of what he calls "abnormal instances" seems less than persuasive.[18] For it consists in assimilating such instances to erroneous judgements generally, while insisting that fallibility is our lot. Should we discover that the counterfactual is falsified in a deviant way, we should admit our mistake and replace the antecedent by a stronger one, stipulating away what

[14] A similar objection is made in Ichikawa and Jarvis (2013, p. 203). I have encountered the suggestion that Malmgren might instead construe an outcome disagreement as "$\Diamond\exists x\exists p$ (GC(x,p) & JTB(x,p) & K(x,p))," thus shifting the locus of dissent from the entire judgement to the knowledge predicate. But this cannot be right, since it is compatible with what the dissenter is denying, namely ($i_{GM}$).

[15] For a different propositional model also meeting them, see Sorensen (1992). For criticism of Sorensen's model, see Häggqvist (1996).

[16] I am indebted to an anonymous referee for prompting this clarification.

[17] For elaboration and defence of the model, see Häggqvist (1996, 2009).

[18] Ichikawa and Jarvis (2009, p. 226) comment: "Incredibly, he seems willing to bite the bullet on this point, and to admit that in such cases, our thought experiments are defective." Malmgren (2011, p. 280) complains that his reply seems ad hoc, since his own objection against a putative proposal construing the judgement as a strict conditional—$\Box\forall x\forall p$(GCxp $\supset$ (JTBxp & ¬Kxp)—appears to be exactly parallel (cf. Williamson 2007, p. 185).

made for the deviance we discovered (Williamson 2007, pp. 200–201; also see his 2009, pp. 468–469).[19] However, this response bypasses the intended point that some sorts of error are possible on his proposal even though they seem innocent—indeed, not errors—on the part of the judger.

Ichikawa and Jarvis (2009) and Ichikawa (2009) emphasize what they consider to be excessive epistemic demands of Williamson's proposal.[20] As Ichikawa puts it, "Williamson's account renders it much too difficult to know the Gettier intuition [judgement]" (Ichikawa 2009, p. 440). This is only partly due to the risks of falsity incurred by deviant realizations: "even in normal worlds, where the counterfactual [$ii_{GW}$] is true as intended", it remains too difficult, Ichikawa argues, to know (2009, p. 440).

As Ichikawa and Jarvis note, counterfactuals like ($ii_{GW}$) are, on standard semantics such as Lewis (1973) or Stalnaker (1968), contingent (Ichikawa and Jarvis 2009, p. 225). They take this to entail that they are knowable only a posteriori. Against this, they defend what they take to be "the standard view", that intuitive judgements about cases are both necessary and knowable a priori (Ichikawa 2009, p. 223). As Ichikawa puts it, Williamson's proposal renders thought experimental judgements "*too* contingent" (Ichikawa 2009, p. 436, italics in original).

In response to Ichikawa, Williamson notes that "a strict implication entails the corresponding counterfactual, and the latter suffices to validate the passage from the possibility of its antecedent to the possibility of its consequent, while making an epistemically less risky claim" (2009, p. 466). The counterfactual corresponding to ($ii_{GIJ}$) would be

$$g \Box\!\!\rightarrow \exists x \exists p(JTBx,p \,\&\, \neg Kx,p))$$

Assuming some standard semantics, this restrict the claim to the closest *g*-worlds, in contrast with the proposal on offer, which extends the claim to all worlds where the Gettier story is true. Williamson exploits this against Ichikawa and Jarvis:

> In a world in which it is highly abnormal not to have many alternative sources of knowledge for a given belief, Gettier's text may present a fiction in which it is not true that the protagonist has justified true belief without knowledge. If we are in an abnormal pocket of ignorance within such a world, then the Gettier story does not strictly imply that there is justified true belief without knowledge; perhaps it even strictly implies that there is no justified true belief without knowledge (Williamson 2009, p. 466)

This passage is complex, perhaps even problematic.[21] But its gist is clear: the strict conditional ($ii_{GIJ}$) is strong enough to carry epistemic risks. So why opt for it, given that Ichikawa's and Jarvis's complaint against Williamson's counterfactual 3* was that it was too difficult to know? However, as I will explain next, I think that deviance is more problematic than Williamson admits, but that we should not—as Malmgren, Ichikawa, and Jarvis do—blame it on Williamson's choice of connective.

## 6 Connectives, Variables, and Constants

I believe that Williamson is right to insist on the fallibility (and, typically, contingency) that comes with counterfactuals. If a proposal concerning the form of thought experiments is to generalize to areas like political philosophy, normative ethics, or applied ethics, the requirement that judgements about cases must be construed as necessity claims becomes an intolerable straight-jacket; especially if coupled with a demand that they be (somehow) less immune to error because knowable a priori. The same holds for cases in philosophy of biology, philosophy of chemistry, and so on.

Of course, one may insist that the demands are not supposed to be met by thought experiments in these areas, but should be upheld for e.g. epistemology, and that "traditional" expectations of what philosophy is carry enough weight to motivate special treatment of thought experiments in certain areas. But this seems dubious, as well as premature. All parties supposedly agree that there is some level of generality at which interesting commonalities among (negative) thought experiments may be captured. And the differences between different areas of philosophy where negative thought experiments are used do not appear to be so great as to prevent general formal treatment.

Moreover, as e.g. Williamson (2007, pp. 181–182) notes, negative thought experiments are—just like ordinary experiments—offered as particular *instances* of a theory they are held to conflict with. Their vignettes describe particular scenarios; the verdicts about these are particular,

[19] Grundmann and Horvath (2013) suggest that deviance may be avoided by stipulating that the subject does have justified true belief in the target proposition, and lacks knowledge of it. Whether this move is feasible is discussed in Ichikawa (2009, pp. 441–442) and Malmgren (2011, pp. 287–289).

[20] They also raise the semantic objection.

[21] Just what are we asked to envisage? For ($ii_{GIJ}$) to be false, some *g*-world accessible from the world of the thinker or judger must lack instances of JTB without K. If there is such a world, this is relevant; but whether this holds at the world where the story is told seems immaterial. Also, "the protagonist" seems to not really belong to the content of ($ii_{GIJ}$)—at least not its consequent (cf. Malmgren 2011, p. 306).

too. When we make a judgement about the hypothetical case, we judge what *would* be the case if *that* scenario were to hold. We do not, it seems, make a stronger claim about what *must* be the case.

These considerations speak in favour of the counterfactual. But as I said, I also believe that Malmgren, Ichikawa and Jarvis are right to insist that the deviance complaint goes beyond whatever problems may come with fallibility in general, and hence that Williamson's appeal to fallibility is insufficient as a reply here. Consequently, I think that Ichikawa and Jarvis are wrong to locate the source of deviance in the counterfactual.

A more plausible culprit, I submit, is the choice of representing apparent proper names in vignettes by bound variables. Deviant realizations of ($ii_{GW}$) are, in effect, particular counterexamples to its consequent, which is a universal sentence.[22] But there is something odd in the root idea of construing standard judgements about cases as statements admitting of particular counterexamples, given that these judgements themselves are naturally taken to describe particular counterexamples—just as in the parallel case of an empirical falsification in science. A particular falsifying instance to an empirical theory is not *realized* by various situations or events: it just is one of these situations or events.[23] The radical idea that modal counterinstances are different in this respect seems to be widely accepted, but should be avoided insofar as we wish to understand thought experiments (of the sort at issue) as a modal counterpart to theory testing generally.

Thus if we could represent case judgements (and the other premises in the formalization of a thought experiment) using constants instead of bound variables, it seems that this might serve two objectives. First, it would get rid of one source deviance besetting accounts like Williamson's. Second, it would render particular statements more faithfully.

Could we use constants? For illustration, consider another specific case (adapted from Ichikawa 2009, p. 437):

(Clock) At 8:28, Smith looked at a clock to see what time it was. The clock was broken; it had stopped exactly 24 h earlier. Smith believed, on the basis of the clock's reading, that it was 8:28

Let "s" denote the fictional Smith of (Clock), and let "a" denote the proposition (or better: proposition schema) that the time was 8:28. Then the judgement that Smith, while justifiedly and truly believing that the time was 8:28, would not know this in the scenario. Formally [with "Cl(x,p)" as the predicates satisfied just by persons and propositions related exactly as in (Clock)]:

($ii_{GH}$) $Cl(s,a) \mathrel{\Box\!\!\rightarrow} (JTB(s,a) \,\&\, \neg K(s,a))$.

This keeps the counterfactual. It also retains, as seems proper, the contingency that comes with a counterfactual plus whatever fallibility that carries. Just as its counterpart statements in a formalization of an ordinary experiment, it is particular rather than general.[24] And it gets rid of intuitively irrelevant "realizations" of the judgement by various possible or actual people, besides the intended protagonist, happening to satisfy the predicate "Cl", since the antecedent now concerns the protagonist introduced in the vignette, rather than an existential or universal generalization, as in the three proposals mentioned so far.

How much does this help? Of course, assuming a Lewis semantics, there is still implicit generalization over the closest antecedent-worlds whenever a counterfactual is asserted. If in some of these, Smith knew in advance that the clock he looked at had stopped exactly 24 h earlier, the counterfactual would be false (to vary the deviance complaint raised by Ichikawa (2009, p. 437) against Williamson). But it is plausible that such worlds are not as close as those rendering ($ii_{GH}$) true. If they aren't, these deviant worlds will not matter for the counterfactual.

On the other hand, of course, if such worlds *are* among the closest antecedent-worlds, ($ii_{GH}$) is false.[25] Nothing explicitly stated in (Clock) immunizes it against falsity; neither, I think, do whatever conventions we may share for contemplating fiction (to which Ichikawa and Jarvis appeal); nor do authorial speaker intentions on the part of the case inventor seem sufficient to guarantee the counterfactual's truth. But the issue at hand is whether such falsity should count as deviance. To object to ($ii_{GH}$) as a rendering of the orthodox judgement about (Clock) solely on the grounds that it risks falsity risks begging the question against a fallibilist about case judgements. Recall that we seek a general account: not even a rationalist would want to embrace infallibilism about thought experiments.

What we get rid of, then, is deviance stemming from odd, unintended individuals who happen to satisfy the case: making it this particular achieves at least that.[26] It is

[22] Ichikawa (2009, p. 442) explicitly calls them "counterexamples to the content of the Gettier intuition [construed as ($ii_{GW}$)]".

[23] Of course, higher-lever generalizations (about laws, say) may have counterinstances that are general and admit of instances (e.g. some law). But the relation between theories criticized via negative thought experiments in philosophy and judgement about such thought experiments is not happily modeled on such relations.

[24] A caveat to this will be issued in the next paragraph.

[25] I am grateful to an anonymous referee for pressing this point.

[26] Thanks to Manuel Garcia-Carpintero for prompting this clarification.

perhaps worth noting that this sort of deviance is clearly what Malmgren (2011, p. 279) discusses by way of her unintended (and fictive) "uncle Joe"; what Ichikawa and Jarvis (2009, p. 226) discuss when they note that someone, unbeknownst to the case inventor, may happen to satisfy the text of a vignette deviantly; what motivates Williamson (2007, pp. 200–201) to contemplate domain restriction; and what occupies Malmgren (2011, p. 306) when she objects to Ichikawa and Jarvis on the grounds that their ($ii_{GIJ}$) risks becoming *true* in a deviant way, since the subject(s) satisfying the consequent of the strict conditional "need not be the same subject as the subject who plays the 'Smith role'" in a world where the fiction established by (S) is true (on their account).

I should perhaps emphasize that I am not here proposing that we assimilate thought experiments to fiction generally, or apply some particular semantics or metaphysics for fiction in general to vignettes.[27] I don't know what the right account of truth in fiction is, and don't have a stance on whether there are important semantic differences between vignettes used in thought experiments (in the current sense) and ordinary narrative fiction, as claimed for instance by Malmgren (2011). The metaphysics could fall any which way compatible with some way for an antecedent like "Cl(s,a)" to be possibly true. What I am mainly concerned with is that "s" not be treated as a mere bound variable.

In any case, fictionality isn't exactly the issue here. Thought experiments certainly often employ straightforwardly invented names, and typically (what with the brevity of many vignettes), there is little by way of explicit stipulation for a contemplator of a case to draw on for forming an image of (or mental file for) these fictional characters. But thought experiments may also use referring terms, sometimes for protagonists. In presenting Newcomb's problem, Nozick (1969) uses "you".[28] For the Chinese Room, Searle (1980) uses "I". And in presenting various cases in *Reasons and Persons*, Parfit (1984) uses both, as well as various fictional proper names. In discussing divergence miracles, Lewis (1981) invokes a hypothetical case involving, it would seem, Richard M. Nixon. It seems natural to construe these apparent names simply as names. And it seems to make little difference whether a case uses prima facie referring or non-referring names of protagonists (or other denizens of their scenarios). Hence it seems natural, and desirable if we want a proposal to generalize, to construe judgements about cases as operating with singular terms irrespective of whether these refer or not in the actual world.

Moreover, the very same sort of counterfactual thinking would seem to take place whether we consider a non-actual possibility involving an actual subject or a merely invented one: there is clearly no "imaginative resistance" stemming from having to contemplate an invented subject, compared to contemplation of existing ones.[29] Whether or not vignettes count as fiction, in a sense covered by any metaphysical account of fiction, and their protagonists as ficta, we clearly have an ability to understand and think about them which doesn't seem to be sharply separable from hypothetical thinking in general. Together, these considerations seem to me additional reason for preferring constants in a QML proposal.

Let us assume that the premises

$(i_{GH})$ $\quad \Diamond GC(s,a),$

$(ii_{GH})$ $\quad GC(s,a) \mathrel{\Box\!\!\rightarrow} (JTB(s,a) \,\&\, \neg K(s,a))$

are true. How is this a counterexample to the JTB thesis, as intended? As per the propositional proposal I sketched above (Sect. 4), I think the thought experimenter is committed to a further premise, which might be articulated as the material conditional

$(iii_{GH})$ $\quad K{=}JTB \supset GC(s,a) \mathrel{\Box\!\!\rightarrow} (JTB(s,a) \,\&\, K(s,a)).$

Hence, the main argument to the conclusion that the JTB theory is false is valid.

As I said, I am unconvinced that a quantified model will be motivated in most contexts for someone interested in thought experiments in general while adhering to Quine's Maxim of Shallow Analysis. But for some purposes such a model may be welcome.

**Acknowledgements** I thank Anna-Sara Malmgren, Daniel Cohnitz, and Mike Stuart for valuable discussion of the ideas I have tried to put forward here. I am also indebted to the editors and anonyous referees of TOPOI for very valuable comments, and to audiences at the University of Urbino and the University of Macau, where parts of this work were presented. I acknowledge support from the Swedish Science Council (grant 421-2012-1004).



[27] Thanks to an anonymous referee for pressing me on this.

[28] Some research suggests that people tend to understand themselves to be the protagonists of vignettes using the second person singular (Tobia et al. 2013). This seems compatible with such vignettes being fictional (and understood to be so).

[29] Note that this is not the often-made point (e.g., Williamson 2005), that an actual *case* would do as well as a hypothetical one for refuting a philosophical claim.

## References


Chudnoff E (2011) The nature of intuitive justification. Philos Stud 153:313–333

Cohnitz D (2006) Gedankenexperimente in der Philosophie. Mentis, Paderborn

Dennett D (1987) The intentional stance. The MIT Press, Cambridge, MA

Gettier E (1963) Is justified true belief knowledge? Analysis 23:121–123

Grundmann T, Horvath J (2013) Thought experiments and the problem of deviant realizations. Philos Stud 170:525–533

Häggqvist S (1996) Thought experiments in philosophy. Almqvist & Wiksell International, Stockholm

Häggqvist S (2009) A model for thought experiments. Can J Philos 39:55–76

Hare R (1984) Moral Thinking. OUP, Oxford

Horowitz T, Massey G 1991: Thought experiments in science and philosophy. Rowman & Littlefield, Savage

Ichikawa J (2009) Knowing the intuition and knowing the counterfactual. Philos Stud 145:435–443

Ichikawa J, Jarvis B (2009) Thought-experiment intuitions and truth in fiction. Philos Stud 142:221–246

Ichikawa J, Jarvis B (2013) The rules of thought. OUP, Oxford

Kühne U (2005) Die Methode des Gedankenexperiments. Suhrkamp, Frankfurt

Lewis D (1973) Counterfactuals. Blackwell, Oxford

Lewis D (1981) Are we free to break the laws? Theoria 47:113–121

Malmgren AS (2011) Rationalism and the content of intuitive judgements. Mind 120:265–327

Nagel J (2013) Intuitions and experiments: a defense of the case method in epistemology. Philos Phenomenol Res 85:495–527

Nozick R (1969) Newcomb's problem and two principles of choice. In: Rescher N (ed) Essays in honour of Carl G. Hempel. Reidel, Dordrecht, pp 114–146

Parfit D (1984) Reasons and persons. Clarendon Press, Oxford

Pust J (2016) Intuitions. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy (Spring 2016 Edition). Stanford University, Stanford. http://plato.stanford.edu/archives/spr2016/entries/intuition/

Putnam H (1975) The meaning of "Meaning". In: Mind, language and reality: philosophical papers, Vol 2. Cambridge University Press, Cambridge

Putnam H (2015): 'Reply to Ian Hacking'. In Auxier R (ed) The philosophy of Hilary Putnam. Open Court Press, LaSalle

Searle J (1980) Minds, brains, and programs. Behav Brain Sci 3:417–424

Sorensen R (1992) Thought experiments. Oxford University Press, Oxford

Stalnaker R (1968) A theory of conditionals. American philosophical quarterly monographs 2 (Studies in logical theory), pp 98–112

Strevens M (2012) The explanatory role of irreducible properties. Noûs 46:754–780

Thomson JJ (1971) A defense of abortion. Philos Public Affairs 1:47–66

Thomson JJ (1973) Killing, letting die, and the trolley problem. The Monist 59:204–217

Thomson JJ (1985) The trolley problem. Yale Law J 94:1395–1415

Tobia K, Buckwalter W, Stich S (2013) Moral intuitions: are philosophers experts? Philos Psychol 26:252–266

van Quine WO (1960) Word and object. MIT Press, Cambridge

Weatherson B (2003) What good are counterexamples? Philos Stud 115:1–31

Weinberg J, Nichols S, Stich S (2001) Normativity and epistemic intuitions. Philos Top 29:429–460

Williamson T (2005) Armchair philosophy, metaphysical modality and counterfactual thinking. Proc Aristot Soc 105:1–23

Williamson T (2007) The philosophy of philosophy. Blackwell, Oxford

Williamson T (2009) Replies to Ichikawa, Martin and Weinberg. Philos Stud 145:465–476