Philosophical expertise put to the test

Abstract

The so-called expertise defense has it that philosophers have skills superior to lay subjects when it comes to making judgements in philosophical thought experiments. Although the nature of philosophical expertise (should it exist) is controversial, it makes for an empirically testable hypothesis: philosophical education and training should improve judgements in philosophical thought experiments. In this paper, we tested this hypothesis on the basis of three skills which we identify as crucial for thought experimentation. We found that philosophers do indeed have better skills in thought experimentation than the folk.

1 Introduction

Experimental philosophers have found that judgements in philosophical cases, or *case judgments* for short, vary with all kinds of prima facie extraneous factors, such as socioeconomic status, gender, order of presentation, cultural background, and more (Alexander and Weinberg 2007, Machery 2017). From these results, some have concluded that case judgements and the *method of cases*, namely the method of using judgments in thought experiments as evidence in philosophical theorizing, is unreliable (Alexander and Weinberg 2007, Machery 2017). Critically, much of the claim about the *general* unreliability of case judgements is based on an induction from experiments with subjects with little or no philosophical training, also known as 'the folk'. Champions of the traditional philosophical methodology have accordingly sought to block this inductive inference by arguing that the folk are not as competent as philosophers are in making case judgements. More specifically, they argue that philosophers' judgements, in virtue of philosophers' professional expertise, are less prone to vary with extraneous effects. This reply is known as the *expertise defense* and it is thought by many to be the (at least prima facie) most promising way of reacting to the challenge posed by experimental philosophers (Devitt 2011, Hales 2006, Horvath 2010, Ludwig 2007, Williamson 2007, 2011).

There have been arguments about the *burden of proof*: experimental philosophers have argued that philosophers must demonstrate that they are not subject to the same kinds of extraneous effects as the folk (Alexander and Weinberg 2014, Weinberg et al.

2010), and defenders of the method of cases have responded that the default assumption should be that professional philosophers' case judgements are more reliable than those who have no or little training in philosophy (Devitt 2011, 2012a, Williamson 2011). Armchair critics, in turn, do not deny that philosophers have *some* expertise, such as "clarifying concepts", "drawing distinctions", and "assessing arguments", but they do question the idea that philosophers have any specific expertise when it comes to making judgements in thought experiments (Machery 2017, 160, cf. Weinberg 2010).

There is some experimental work which has directly tested the reliability of case judgements by philosophers -- without having to rely on any inductive inferences from the folk to the 'experts' (Feltz and Cokely 2012, Schulz et al. 2011, Schwitzgebel and Cushman 2012, 2015, Tobia et al. 2013, Wiegmann et al. forthcoming). For example, Schwitzgebel and Cushman (2012, 2015) have shown that even philosophers are subject to order-effects in the famous trolley-footbridge cases. Most of this work, however, has been done in moral philosophy, and it is not clear how far it extends to other realms of philosophy. We are aware of only two studies outside the moral realm, both of which have mixed results (Horvath and Wiegmann 2015, Machery 2012).

Machery (2012) found that although the majority of subjects reported standard Kripkean intuitions regarding the reference of proper names, the proportion varied across groups (66.7% to 88.6%). In particular, philosophers and semanticists were more likely to report Kripkean intuitions than linguists with other specialisations. Machery takes this to be evidence for intuitions to be biased by the particular theories to which practitioners of certain academic fields have been exposed. Yet, contrary to what Machery claims, the results are also compatible with the expertise defense: philosophers and semanticists are simply better trained than other experts in making reliable judgements about reference (Devitt 2012b, 23-24).

Horvath and Wiegmann (2015) studied the responses of professional philosophers with a specialization in epistemology to the responses of lay subjects with regard to variations of fake-barn scenarios. Even though Horvath and Wiegmann found that their polled experts in the majority of cases agreed with the 'textbook consensus' and the folk didn't (in agreement with the expertise defense), they also obtained some unexpected results. In the fake-barn cases, they found that even though experts' responses were on average closer to the disagree end of the scale than the folk with regards to knowledge ascriptions (as expected), more experts agreed than disagreed with the textbook consensus (contrary to expectations). Even more surprisingly, perhaps, they found that laypeople ascribed knowledge to clear non-knowledge cases.

Thus, even if one may accept that there is direct evidence that philosophers themselves are subject to extraneous effects in the moral realm, one may be more skeptical about the experimentalists' challenge in other realms of philosophy, where the challenge depends more strongly on an inductive inference from the unreliability of case judgements by the folk. In fact, different case judgements may be generated by different mental processes and we therefore shouldn't presume that all kinds of case judgements are equally reliable (Nado 2014). The expertise defense is thus still a prima facie reasonable reaction by the methodological traditionalists to the experimental challenge.

In this paper, our guiding idea was that if philosophers really do have an expertise in thought experimentation, then they should have it in virtue of year-long formal education, supervised learning, and year-long experience.[1] This makes for a testable hypothesis: philosophers should be better at thought experimentation than the folk *by virtue of their training in philosophy*.

Here is how we will proceed: In Section 2, we discuss extant views of philosophical expertise and present the three skills and argue that they are necessary for good thought experimentation. In Section 3, we present our hypotheses and our methods. In Section 5 we present our results. Section 6 concludes this paper.

2 What is philosophical expertise and how can it be tested?

Defenders of 'armchair philosophy' argue that philosophers have an expertise in making judgements in thought experiments. But what precisely is this philosophical expertise based on? A common view amongst proponents of the expertise defense is that it has to do with the sophisticated application of concepts. For example, Williamson (2011, 191) writes that philosophers "apply general concepts to specific cases with careful attention to the relevant details". Likewise, Ludwig (2007, 138) believes that professional philosophers have honed "skills in responding to questions about described scenarios on the basis of one's competence in concepts involved". Horvath (2010) details that philosophers have "a

[1] Research in the psychology of expertise suggests that expertise in other domains is often acquired gradually in the course of education and professional training (e.g. Chi et al. 2014, Ericsson 2006).

superior ability to apply the relevant concepts, like KNOWLEDGE or FREE WILL, to a wide range of actual and hypothetical cases" because they have developed 'conceptual sensitivity' and 'interpretive ability', which he cashes out as philosophers being "more sensitive to potential ambiguities, unclarities or incoherencies in applying these concepts" and as philosophers being able to "rule out a lot of philosophically unintended interpretations", respectively. With regard to expertise required in thought experimentation in particular, Williamson (2011) summarizes: "one must read and digest the description of the scenario … then one must judge what would be the case in the scenario described … one must also judge whether the scenario is really possible … finally, one must determine whether the premises, if verified, do entail the proposed conclusion" (224).

It is useful to distinguish between two models of expertise: the 'mastery model' and the 'thought experimentation model' (Machery 2017, 158ff.). According to the mastery model, philosophers have a better 'grasp' of philosophical concepts in that they are better than the folk at applying philosophical concepts such as truth, knowledge, necessity, etc. to their occurrences. According to the thought experimentation model, philosophers are better than the folk at "ignoring the irrelevant (e.g., narrative) aspects of thought experiments and at singling out their pertinent features" and they are also better at "understanding how a concept applies to the type of unusual situation described by a thought experiment" (ibid., 158; cf. Weinberg et al. (2010, 348)). In this paper, we are most interested in the thought experimentation model of philosophical expertise, of which Machery makes rather short shrift:

> Graduate students in philosophy are not taught systematically how to thought experiment, what kinds of feature to ignore in general, what kinds of feature to pay attention to in general, etc. (At best, they are told to ignore or consider particular features in specific thought experiments.) There are no textbooks for thought experimenting or graduate seminars teaching how to thought experiment. Thought experiments are often written in a way that invites errors, suggesting that philosophers have not thought hard about what a good thought experiment is or about how to write one. In particular, thought experiments often contain philosophically irrelevant narrative elements, which may have a distorting influence on judgments. (Machery 2017, 165)

Even if it were true that there were no explicit, systematic, and general instructions in our teaching of how to carry out thought experiments, there could of course still be implicit learning acquired on the basis of examples. For example, students can implicitly pick up on important features of thought experiments by simply being exposed to paradigms and variants of it. The argumentative context in which a thought experiment figures can also give cues about which feature of a thought experiment are good features. For example, given the impact Gettier cases have had on philosophical thinking about knowledge, and given that this is being communicated to students, students can infer that some features of Gettier cases are features that characterize good thought experiments. For example, Gettier cases satisfy all the three conditions that have traditionally been associated with knowledge, and yet, exhibit no knowledge. Students can thus infer that good thought experiments provide counterexamples to the conditions laid out by the theories they target. That of course doesn't mean that all good thought experiments need to provide counterexamples (even though many do; see Häggqvist (2009)).

If the thought experimentation model of philosophical expertise were correct and philosophers were better trained in thought experimentation than the folk (whether explicitly or implicitly), then there should be *skills* associated with thought experimentation that can be practiced and honed. What are candidates for such skills?

2.1 Three skills of thought experimentation

We take it that good skills of thought experimentation should make case judgements *informed* and *informative. Informed* judgements are those that subjects make on the basis of information that is necessary for grasping and evaluating the scenario of the thought experiment. *Informative* judgements are those informed judgements that can be put to further use in philosophical debates. We identified three skills which we believe satisfy this description.

First, most basically and most importantly, one must be able to comprehend the vignette describing the thought experiment: if one doesn't understand the scenario described in the vignette, one cannot make an informed case judgement about this scenario. More specifically, one must be able to understand what Machery (2017) has called the 'target content' of a thought experiment: "facts and non-factual connotations that matter from a philosophical point of view" (13). Machery distinguishes the target content from the superficial content, which concerns "the narrative setting of the philosophical case" and consists of "facts and non-factual (e.g., emotional) connotations that do not matter from a philosophical point of view; it varies across versions of a given case" (13). For example, the superficial content of Gettier cases is very dissimilar: some involve malfunctioning clocks that accidentally indicate the correct time, others the misidentification of who in one's office in fact owns a Ford. The target content of Gettier cases, on the other hand, always involves an agent who has justified true beliefs that X by luck. It is this content one must comprehend in each single thought experiment in order to make an informed case judgment.

For example, in Gettier cases, subjects must understand that Smith has justified true belief in order for their judgements to be informed judgements about whether or not Smith knows. If subjects were to understand that Smith has justified belief but not that he has *true* justified belief, or if they were to understand that Smith has true belief, but not that he has true *justified* belief, then subjects' judgements wouldn't be informed and therefore not as valuable as the judgements by subjects who fully comprehend the target content. And of course subjects must understand that Smith has justified true belief only by luck. Likewise, in the Gödel-Schmidt cases, subjects must understand that Gödel, contrary to actual fact, is stipulated not to have proven the incompleteness of arithmetic. Subjects who do not understand this element of the scenario can still make a judgement about whether the name "Gödel" refers to the individual Gödel, but their judgements will not be as informative as the judgements by subjects who do understand this presupposition. Hence, the first skill we identify for successful thought experimentation is:

(1) The ability to comprehend the target content of a thought experiment.

The second skill we identify is closely related to the first skill, namely:

(2) The ability to distinguish between the target content of a thought experiment from the superficial content.

Again, this ability allows subjects to extract the target content not only across superficially varying cases (as suggested by Machery's definition), but also in individual cases. For example, in a Gettier case, in which Smith has justified true belief but no knowledge that somebody in his office owns a Ford on the basis of seeing Jones drive one (even though it is not Jones but his colleague Brown who owns a Ford), this ability should allow one to understand the fact that Jones rents a Ford (rather than occasionally drives Brown's whenever Brown is sick) is irrelevant to grasping the target content. Likewise, this ability should allow one to understand that it is quite irrelevant to the target content that the main character of Jackson's famous thought experiment is a woman called Mary or a man called Martin and whether the main character knows “all the physical information there is to obtain about what goes on when we see ripe tomatoes” or whether she knows all the physical information there is to obtain about what goes on when we see red cars (Jackson 1982). The ability to discern the target content from the superficial content of thought experiments is important for successful thought experimentation because without this ability, there is a risk one doesn’t see the forest for its trees, i.e., one risks getting sidetracked when considering one’s intuitions regarding the case in question.

The hypothesis that the experts should do better than the folk with regards to both the discerning of the target content from the superficial content and the comprehension of the target content can be motivated by exploiting a point made by Machery in his criticism of the method of cases. According to Machery, the method of cases is unreliable, because thought experiments exhibit three characteristics that are “disturbing” to subjects, causing the demographic and presentation variables to influence case judgements negatively (112ff.). First, thought experiments are “unusual”. For Machery, a thought experiment is unusual “if and only if we encounter it infrequently or if we rarely read texts about it” (113). This unusual nature, according to Machery, can be caused by either its superficial content or its target content (as defined above). Twin earth, for example, is unusual in its target content and the fake barn case is unusual also for its superficial content (113-114). The superficial content of a thought experiment can be particularly confusing to subjects because contains many irrelevant details, are rich in narrative content, and “describe hypothetical and actual situations in a vivid and lengthy manner” (119). Machery also claims that “it is unlikely that the target content and the superficial content can be fully disentangled” by subjects engaging in thought experimentation (ibid.). Lastly, thought experiments “pull apart what usually goes together”, for example, footbridge cases “pull apart” the usual association of physical violence and “doing more harm than good” and Gettier cases pull apart the usual association of knowledge with true justified beliefs (116). Because of these three disturbing characteristics of thought experiments, Machery (2017, 112) deems philosophical cases outside the ‘proper domain’ of reliable judgements. Relying on a very substantive inductive inference from findings in the folk to the ability of philosophers, Machery takes these three factors to be a reasons for mistrusting and for abandoning the method of cases altogether.

Even if Machery is right about the disturbing effect of the three characteristics he identifies, the strength of the effect need not be equal for all subjects. In fact, Machery himself concedes that the proper domain of reliable judgements "varies with the expertise of the person judging" (112). This opens the possibility that some subjects are better suited to make certain judgements, namely those who possess an expertise in the domain in question. If philosophers have an expertise in thought experimentation, one would expect their judgments to be more reliable in that domain, even if they too are affected – to a lesser degree – by the disturbing characteristics of thought experiments. Indeed, thought experiments are certainly more frequent for philosophers than they are for the folk (and therefore more "usual" in Machery's sense). Accordingly, philosophers are much more used to the sometimes outrageous superficial content of thought experiments and the pondering of scenarios in which properties are pulled apart. And they certainly know much better "what matters from a philosophical point of view" when discerning the target content from the superficial content of thought experiments. So if Machery is right about the disturbing effects of thought experiments, then it would seem – contrary to what Machery wants to argue – that philosophers should be better subjects than the folk in thought experimentation. But whether they really are should be tested experimentally.

A third ability is briefly mentioned by Williamson in the aforementioned quote: "one must also judge whether the scenario is really possible, for otherwise the thought experiment may not fit the purpose" of refuting a philosophical view, one may add (Williamson 2011, 224). Williamson's brief remark can be unpacked as follows: thought experimentation requires the willingness to accept (for the purposes of the thought experiment) as possible the scenarios described in thought experiments. For example, if one does not accept that the watery substance on twin earth is a metaphysical possibility, and if one insists that substances that possess all the 'superficial' qualities that water does must be H_2O (as in the actual world), then one will not be in a position to judge whether or not people on twin earth refer to water when they talk about the watery substance on twin earth. As Machery (2017, 162) points out, it is important that 'budding philosophers' are made aware of the fact that twin earth is not meant to be actual. Likewise, in thought experiments about personal identity which involve brain-swapping, one must accept the metaphysical possibility of such a process, even though such a process may well be physically impossible. Again, not accepting such metaphysical possibilities (for the purpose of the thought experiment) will not allow one to make a judgement (either way) whether personal identity requires identity in outward appearance. It is also worth noting that the acceptance of the relevant possibility premise is necessary in Williamson's counterfactual model of thought experimentation (Williamson 2007, 184). Thus, the third ability essential for thought experimentation is:

(3) The ability to accept as (metaphysically) possible the scenarios described in the vignette.

It may well be that naïve subjects are less willing to accept such scenarios when they are physically impossible or highly unlikely than the philosophers are. In fact, given what we pointed out above already about the unusualness of thought experiments, there is reason to think that philosophers are more inclined to accept the metaphysical possibility of thought experiments than the folk. More specifically, one could argue that since thought experiments are unusual (in Machery's sense) to the folk but not to the philosophical experts, philosophers should be more willing to accept a wider range of possible scenarios than the folk.

In sum, we find it plausible that the ability to comprehend the target content of thought experiments, the ability of telling apart the target and the superficial content of thought experiments, and the ability to accept (metaphysical) possibilities are critical for good thought experimenting. Making informed and informative case judgements (in the sense defined above) requires all of these abilities.

2.2 Thought experimentation and the textbook consensus

In order to test case judgements, one must use a standard against which these judgments are measured. This standard is what has been called the 'textbook consensus' of case judgements (Horvath and Wiegmann 2015), TBC for short. For example, it is the TBC that the agent in Gettier cases does not have knowledge of X even though they have justified true belief of X. Even though not many experimental philosophers are explicit about using this standard, they use it all the same. For example, Machery et al. (2004) compare their cross-cultural results in Gödel cases against the standard view of the causal theory of reference and challenge philosophers to reconsider their methods on the basis of this standard being violated in some of their (non-Western) subjects.

Indeed, although some armchair critics do not deny the existence of a TBC about case judgements (Machery 2017, 127-128), they regard it as simply "reinforced" through "intense training and selection" without there being good *epistemic* reasons for the particular case judgements (Machery 2017, 128, Machery et al. 2004, B9, Weinberg et al. 2010). We find this 'sociological' thesis implausible for several reasons.[2] There are few academic disciplines in which critical discourse is exercised to the extent as it is in philosophy. As van Inwagen (2004, 334) once put it, "disagreement in philosophy is pervasive and irresoluble. There is almost no thesis in philosophy about which philosophers agree." In a very well-known recent survey amongst professional philosophers with thirty questions on a whole range of important topics in analytical philosophy, the leading views attracted more than 60% support in only 7 out of 30 questions. In 11 questions the leading view had even only 36% or less support (Bourget and Chalmers 2014, Chalmers 2015). Thus, if there is any sociological reinforcement in philosophy (for no good epistemic reasons), then it doesn't seem to be very successful. In particular, it seems highly unlikely that philosophers are as prone to disagree as they are and at the same time more than happy to blindly and uncritically follow the lead of their teachers (in their education) or leading figures (in their professional careers) *only* when it comes to thought experimentation.

That there is widespread agreement on case judgements despite philosophers' disagreement on many other issues seems to be true. Consider for example the hotly debated Chinese Room thought experiment by Searle (1980). Contrary to the Gettier cases, which are widely agreed to have undermined the view that knowledge is justified true belief, Searle's Chinese Room has given rise to a plethora of views regarding the prospects of strong artificial intelligence. Still, the judgement that the man in the room does not have any understanding of Chinese, despite his successful symbol manipulation, is usually granted by all parties in the debate. For example, those who argue that the entire system understands, grant that the man does not understand Chinese. Likewise, those who insist that a variation of the Chinese Room would have understanding, do not deny that the man in Searle's original thought experiments has no understanding. Even those critical of

[2] Williamson (2011) even calls it a 'conspiracy theory'.

the case judgement, usually do not deny that they would make the same judgement—only that case judgements ought not to be relied upon (e.g. Dennett 1993, Thagard 2014).[3] We therefore do take it that even in the most controversial thought experiments, one can legitimately speak of a textbook consensus about the relevant case judgements.

Still, it may of course be the case that parts of the TBC are mistaken. In order to leave room to this possibility the TBC has been treated as "a defeasible or prima facie standard for assessing the quality of intuitions" which should be given up when experimental results depart too far from the TBC (Horvath and Wiegmann 2015, 2707).[4]

Following Horvath and Wiegmann, we treated the TBC as a defeasible standard. In contrast to Horvath and Wiegmann, however, the TBC was not our only measure of the quality of case judgements: the three skills of thought experimentation we identified provided *independent measures*. Thus, if it could be shown that philosophers not only are more likely to make case judgements concordant with the TBC, but also that they do better in the skill tasks we identified, then they could justifiably be said to make better case judgments than the folk.

3 Hypotheses

In our study, we tested altogether five hypotheses. With our first hypothesis, we sought to test whether the three skills we identified were in fact required for good thought experimentation (as we had speculated). We tested this by asking whether the three skills are actually predictive of the textbook consensus, which we presumed as a defeasible standard of the quality of case judgments.

> ***H1***: *subjects who do well on the three skill tasks are more likely to converge on the textbook consensus in their case judgements than subjects who don't.*

If the three skills are predictive of judgements converging on the TBC, and if philosophers have an expertise in thought experimentation, then the following hypothesis should be true as well:

[3] See for example the standard works by Cole (2015), Heil (2013), and Lowe (2000).

[4] This is what Horvath and Wiegmann end up doing with regards to fake-barn cases, given that the many philosophers they surveyed were willing to assign knowledge. Yet it is noteworthy that they do not give up on standard judgements about clear no-knowledge cases just because the folk assign knowledge.

H2: subjects trained in philosophy are more likely to make case judgements in agreement with the philosophical textbook consensus than subjects who are not trained in philosophy.

If H2 is correct and philosophers are better than the folk in thought experimentation, it should be correct in virtue of philosophers having a better set of thought experimentation skills. Accordingly,

- *H3: subjects trained in philosophy are better at comprehending the target content of thought experiments than the folk, i.e., subjects without training in philosophy.*
- *H4: subjects trained in philosophy are better at discerning the target from the superficial content in thought experiments than the folk.*
- *H5: subjects trained in philosophy are more willing to accept the possibility of thought experiments than the folk.*

4 Methods

4.1 Subjects

In order to test these hypotheses, we tested two groups of subjects: philosophical experts (as defined by holding a PhD or being enrolled in a PhD programme), which we recruited through Philos-L and targeted emails to selected departments, and—in accordance with recent practices in experimental philosophers—a group of non-philosophers recruited through Amazon Mechanical Turk.[5] We first tested our philosophy subject group and then matched the number of subjects we obtained with our survey on MTurk. Of our MTurk subjects we required (via a filter) that they had at least 100 HITs (Human Intelligence Tasks on MTurk) approved prior to our study with an approval rate of more than 97%.[6] Overall, we tested 242 philosophers and 242 non-philosophers. The median age of the group consisting of philosophers and non-philosophers was 35 and 34, respectively. The gender distribution was 176 males and 63 females in the philosophers' group (reflecting the deplorable male-dominance in the field) and 143 males and 99 females in the non-philosophers' group. See Appendix 2 for details about the composition of our subjects.

[5] In the recent large scale study on the reproducibility of studies in experimental philosophy (Cova et al. preprint), it was reported that out of 39 original Xphi studies from 2003-2015, the majority (25) used students. More recent replication studies of these original studies used mostly MTurk workers (29 out of 39).
[6] On MTurk, after a subject completes a HIT, it must be confirmed by the requester that the task was carried out to their satisfaction.

4.2 Materials

To all of our subjects, we presented modified versions (of the superficial content) of a Gettier case, Searle's Chinese Room, Mary's Room, a Fake Barn case, a Gödel case, and Twin Earth. We modified these standard cases in order to minimize recall effects (see Appendix 1 for all used scenarios).

Subjects had four tasks for each thought experiment (the order of the thought experiments we randomized). First, subjects were presented with a statement about the thought experiment and had to indicate their agreement or disagreement (on a 5-point Lickert scale). Half of the statements we presented were correct and the other half incorrect relative to the textbook consensus in order to not bias subjects answers toward either agreement or disagreement. When subjects agreed (disagreed) with what we took to be a correct (incorrect) statement, we counted that as a correct response. Subjects were then presented (in this order) with a task testing their comprehension of the target content, their understanding of relevant/irrelevant detail, and the willingness to accept the possibility of the presented scenario. Each of those tasks was presented on a separate screen. Subjects were reminded of the vignette at the bottom of the screen. To illustrate, consider one of our six thought experiments, namely a version of the Gettier case:

> John is a security guard at a warehouse. One night, the alarm bell rings. John therefore comes to believe that someone has just broken in. Unbeknownst to John, the triggering of the alarm was in fact just the result of some electronic malfunction, because the technician messed up. But as it turns out, there was actually a thief who managed to break into the warehouse at the same time when the alarm rang.

- *Case judgement*: "To what extent do you agree or disagree with the statement 'John does not know that someone has broken into the warehouse'".
- *Comprehension:* "To what extent do you agree or disagree that the following statement is correct? 'John believes that there was a break-in.'"
- *Irrelevant or relevant content*: "The scenario just presented to you contains the following piece of information. To what extent do you agree or disagree that the following information was important relative to that judgement? '…that the technician messed up.' OR 'the triggering of the alarm was in fact the result of some

electronic malfunction'" [subjects were randomly presented with either relevant or irrelevant content]

- *Possibility:* "To what extent do you agree or disagree that the described scenario is possible."

Two clarifications about the presentation of two of our tasks are in order. First, each subject received *either* a statement about the irrelevant content of the vignette *or* a statement about the relevant content. We randomized those statements over the subjects. Second, the statement about the possibility of the task was left vague on purpose. Our rationale was that laypeople might by default employ a narrower notion of possibility than the professional philosophers, who are very much used to ponder counterfactual scenarios and those which are quite far removed from the actual world. Accordingly, our expectation was that the folk would tend to disagree more often with the possibility statement than the philosophers would.

In the following we abbreviate the case judgement, the comprehension task, the relevant/irrelevant task, and the possibility task as JUDGE, COMP, R/IRR, and POSS, respectively.

4.3 Data analysis

In order to test H1, we devised a Negative Binomial Count model. The model predicts whether or not the number of correct answers in each of our three skill tasks can predict the number of JUGDE tasks that we deemed correct (relative to the textbook consensus). In order to test H2, we compared the mean of correctly judged thought experiments by philosophers with the mean of correctly judged thought experiments by non-philosophers. We tested the significance of the difference with a t-test. In order to test H3-H5, we used a t-test to find out whether philosophers would get more COMP, R/IRR, or POSS tasks right across the six thought experiments, respectively, than the non-philosophers. We also used chi-squared tests to determine the significance of the group-differences in the number of correct JUDGE and skills task responses for each thought experiment. Throughout, we distinguished between two ways of counting our subjects' Likert scale responses that we deemed correct (relative to the textbook consensus): a STRICT count (strongly agree/disagree) and a LAX count (strongly and somewhat agree/disagree). Alternatively, we should add, one may have considered comparing the means of *all* responses of the two subject groups (rather than just the 'correct' ones). That, however, would presuppose that the "neither agree nor disagree" response at the mid-point of the Likert-scale is in fact a better response than the lower end of the scale. We see no good justification for such an assumption.

5 Results and discussion

***H1**: subjects who do well on the three skill tasks are more likely to converge on the textbook consensus in their case judgements than subjects who don't.*

We constructed a binomial regression model for each skill for the LAX and STRICT count respectively, for the entire data set. The model for COMP and R/IRR correctly predicted the number of correct JUDGE tasks in both the STRICT and the LAX count (*p=.05*). For POSS, the model predictions were correct for the JUDGE task in the STRICT count only (*p=.05*). We consider H1 to be confirmed in the STRICT count and partially confirmed in the LAX count. Fig. 1-3 depict the results for the LAX count (see also Appendix 3 for further details).

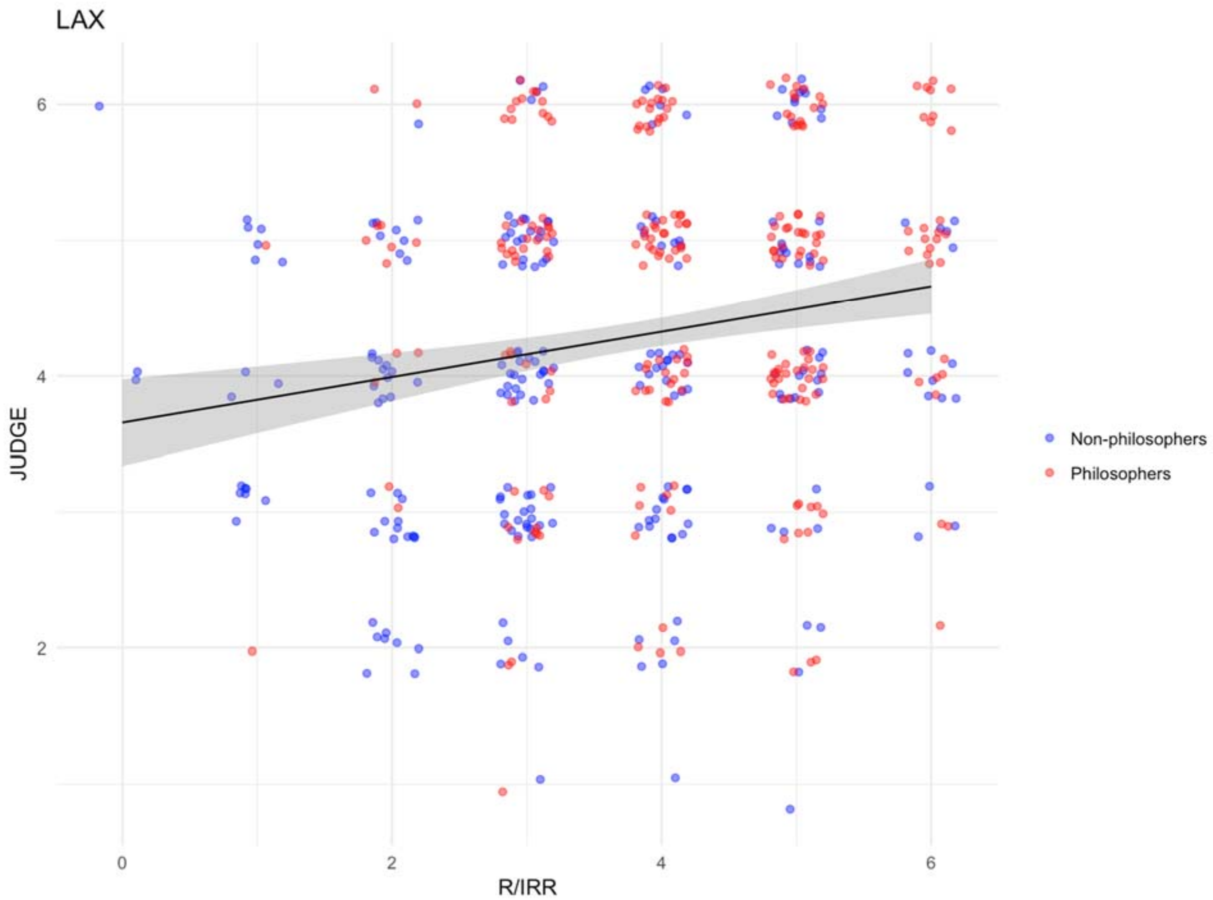

Fig. 1: Performance of all subjects on the R/IRR and JUDGE task. The x- and y-axis represent the number of tasks that subjects got correct on both R/IRR and JUDGE. Blue dots represent non-philosophers and red dots philosophers. There is a clear positive regression.

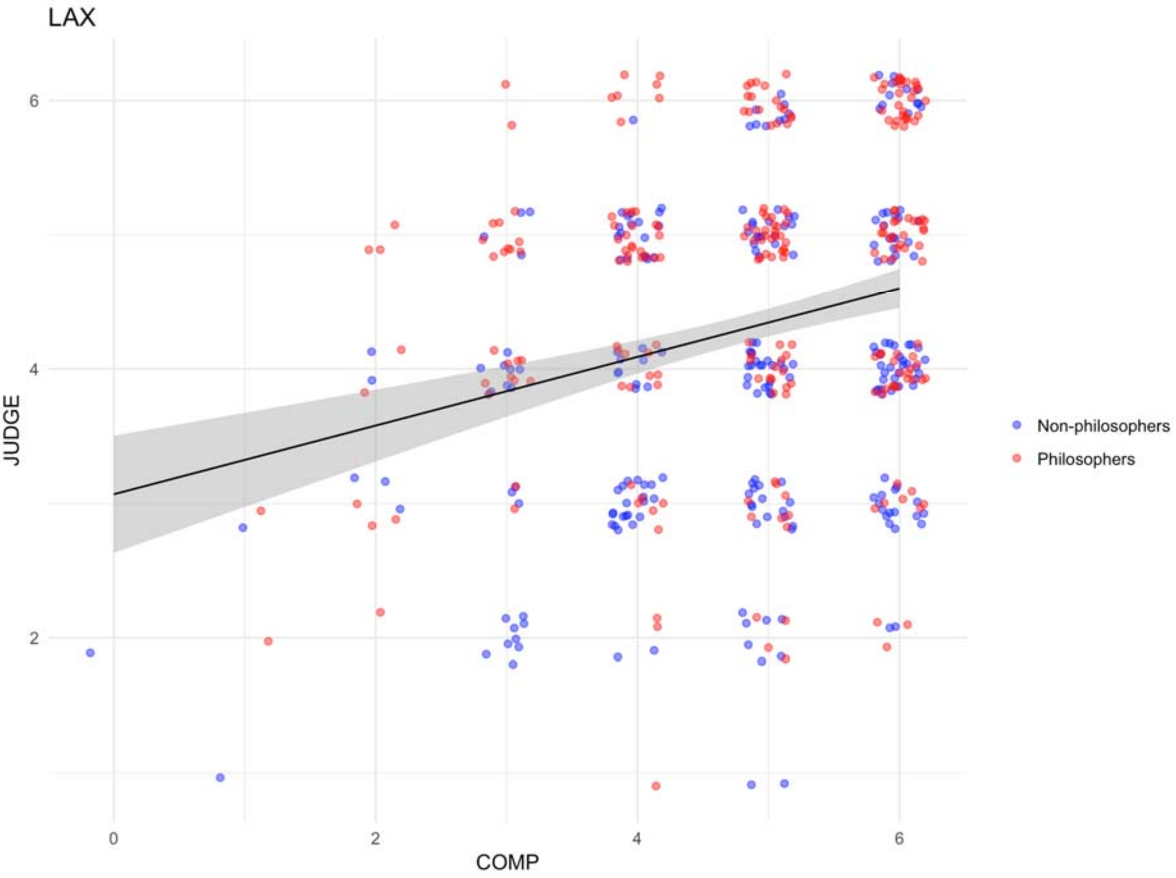

Fig. 2: Performance of all subjects on the COMP and JUDGE task. The x- and y-axis represent the number of tasks that subjects got correct on both COMP and JUDGE. Blue dots represent non-philosophers and red dots philosophers. There is a clear positive regression.

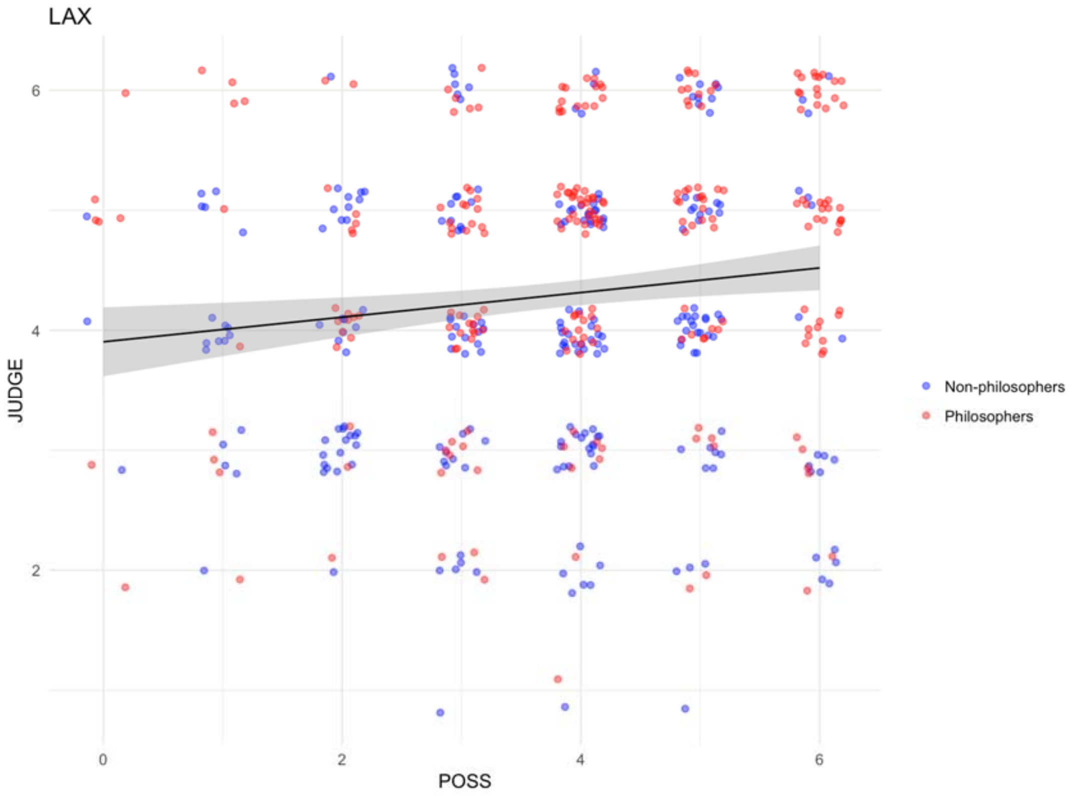

Fig. 3: Performance of all subjects on the POSS and JUDGE task. The x- and y-axis represent the number of tasks that subjects got correct on both POSS and JUDGE. Blue dots represent non-philosophers and red dots philosophers. There is no positive correlation in the LAX count (but there is in the STRICT count).

***H2**: subjects trained in philosophy are more likely to make case judgements in agreement with the philosophical textbook consensus than subjects who are not trained in philosophy.*

Our test revealed a mean of 3.11 of correct JUDGE answers by philosophers across the six tasks in the STRICT count and a mean of 4.62 in the LAX count. Non-philosophers on averaged judged fewer thought experiments correctly (M=2.65 for the STRICT and M=3.98 for the LAX count). The difference in the means of philosophers and non-philosophers was highly significant in both the STRICT *t(480.46)=3.48, p<.001* and in the LAX count *t(481.47)=6.2, p<.001*. Cohens' d is .58 in the STRICT count, and .64 in the LAX count. Thus, H2 is confirmed. Fig. 4 depicts the percentages of correct, incorrect, and neutral responses of our subjects relative to the TBC.

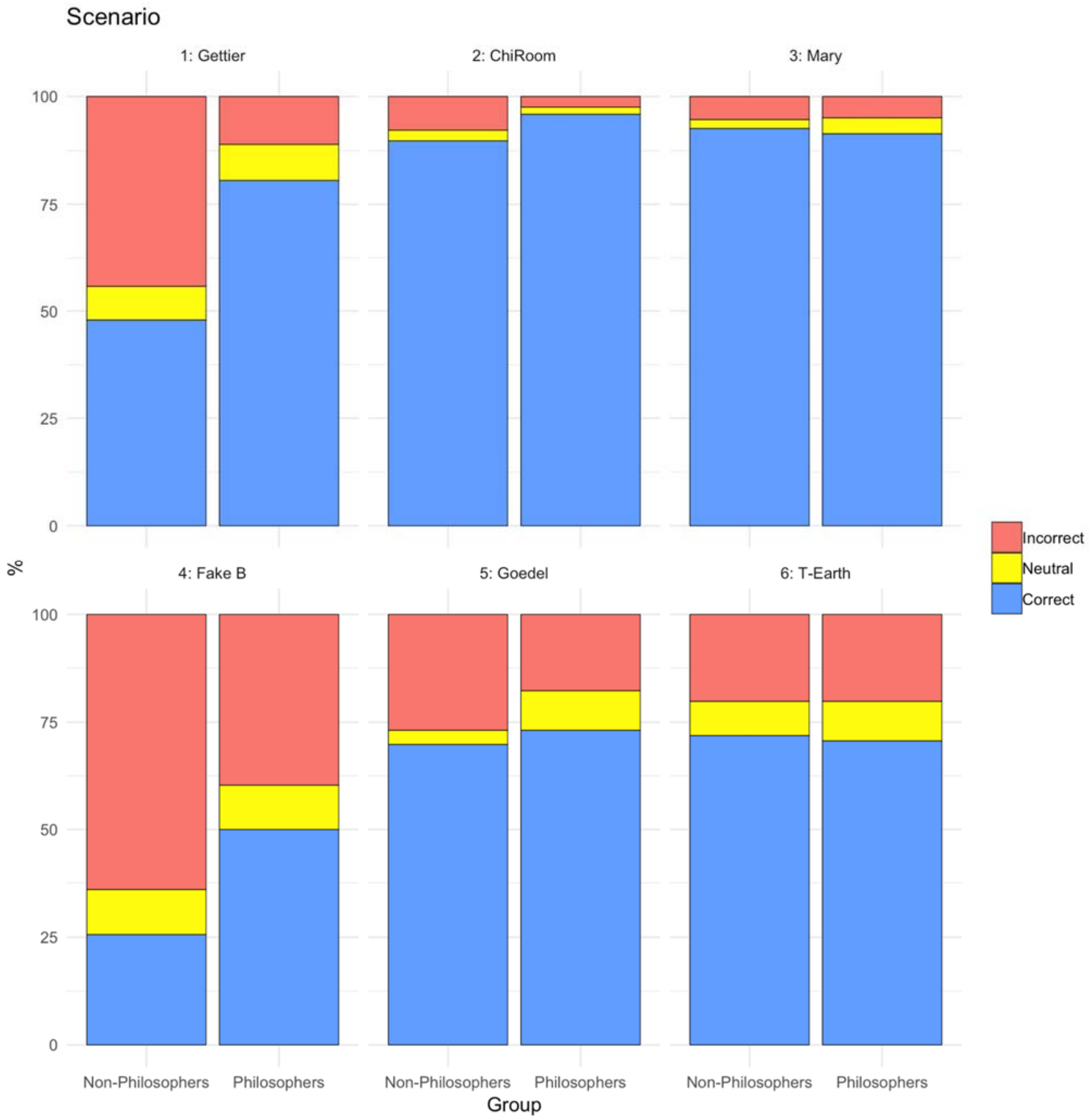

Fig. 4: Results of the JUDGE task in our six thought experiments by both non-philosophers and philosophers (left and right columns, respectively). Blue represents the correct responses (relative to the textbook consensus), red incorrect responses, and yellow neutral responses.

The chi-squared tests we used to determine the significance of the difference between the correct answers by philosophers and non-philosophers, respectively, delivered there following results:

STRICT: Gettier: $\chi^2(1, N = 484) = 46.04$, $p <.05$; Chinese Room: $\chi^2(1, N = 484) = 6.19$, $p <.05$; Mary's Room: $\chi^2(1, N = 484) = 0.09$, *ns;* Fake barn: $\chi^2(1, N = 484) = 16.23$, $p <.05$; Gödel: $\chi^2(1, N = 484) = 0.01$, *ns;* Twin Earth: $\chi^2(1, N = 484) = 0.53$, *ns*.
LAX: Gettier: $\chi^2(1, N = 484) = 56.14$, $p < .01$; Chinese Room: $\chi^2(1, N = 484) = 6.93$, $p <..05$; Mary's Room: $\chi^2(1, N = 484) = 0.25$, *ns;* Fake barn: $\chi^2(1, N = 484) = 30.59$, $p <.01$; Gödel: $\chi^2(1, N = 484) = 0.65$, $p <.05$; Twin Earth: $\chi^2(1, N = 484) = 0.09$, *ns*.
Appendix 3 also lists the results for detailed results, including chi-squared tests for the respective skills tasks.

There are some further aspects of our test of H3 which are also noteworthy. First, although judgments in Gettier cases are widely perceived as the most robust case judgments there are, in our study only 80% of philosophers agreed with the standard judgement, compared to over 90% agreement in Mary's Room and the Chinese Room. Second, the philosophers in our sample agreed least with the TBC with regard to our fake barn case (50%); here the TBC may have to be revised (see Horvath and Wiegmann 2015). Third, the differences between philosophers and the folk was most pronounced in both our Gettier case and the fake barn case: in both cases, philosophers were much less willing than the folk to ascribe knowledge (in accordance with the TBC). That there is such a pronounced difference between the folk and the philosophes *specifically* in epistemological cases is an interesting question for further research.

H3-5*: subjects trained in philosophy are better at comprehending the target content of thought experiments than the folk / better at discerning the target from the superficial content in thought experiments than the folk / and more willing to accept a wider range of possibilities in thought experimentation than the folk.*

Our t-test results are for the STRICT count are as follows:

- R/IRR: *t(477.75)=5.45, p<.05 (philosophers: M=3.2; non-philosophers: M=2.53)*
- COMP: *t(481.87)=1.3, ns (philosophers: M=3.63; non-philosophers: M=3.44)*
- POSS: *t(452.51)=7.48, p<.05 (philosophers: M=2.43; non-philosophers: M=1.35)*

In other words, in STRICT philosophers do better than the non-philosophers with regard to R/IRR and POSS, but not in COMP. Our t-test results are for the LAX count are as follows:

- R/IRR: *t(463.53)=6.85, p<.05 (philosophers: M=4.21; non-philosophers: M=3.43)*
- COMP: *t(481.95)=.23, ns (philosophers: M=4.82; non-philosophers: M=4.79)*
- POSS: *t(479.71)=3.04, p<.05 (philosophers: M=4.01; non-philosophers: M=3.6)*

Thus, just as in the STRICT count, in LAX philosophers do better than the non-philosophers with regard to R/IRR and POSS, but not in COMP. We can therefore conclude that H4 and H5 are confirmed, but H3 isn't. See Fig. 4 for a depiction of the results and Appendix 3 for more detailed results.

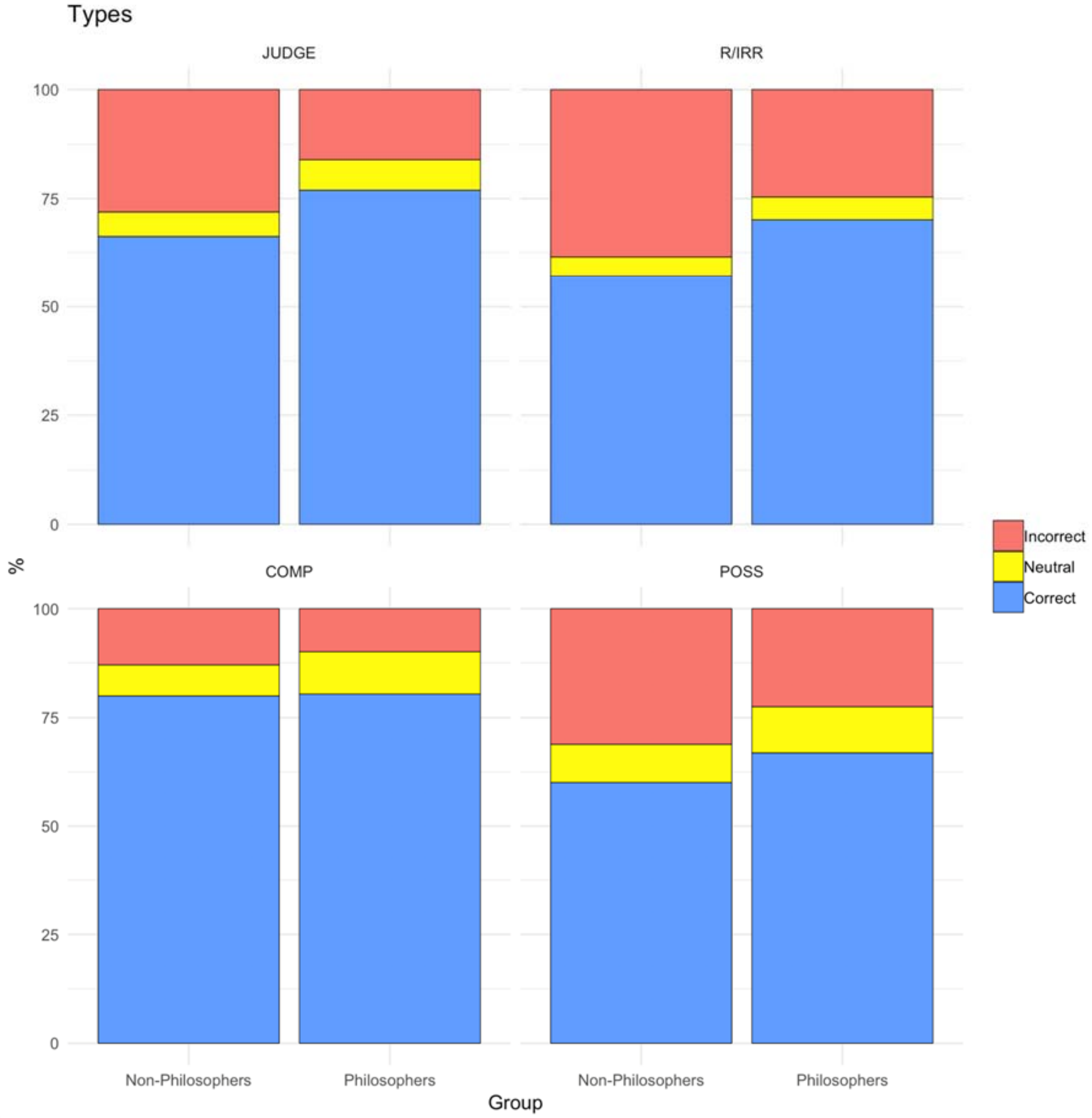

Fig. 5: Results of the JUDGE and the three skills tasks R/IRR, COMP, and POSS for both non-philosophers and philosophers (left and right columns, respectively). Blue represents the correct responses (relative to the textbook consensus), red incorrect responses, and yellow neutral responses.

6 Conclusion

The purpose of this study was to determine whether philosophers possess an expertise when it comes to making case judgments, and in particular, whether they possess superior *trainable skills* of thought experimentation that would allow them to make better case judgments. We identified three skill tasks which we believe are essential for making informed and informative case judgments: the comprehension of the scenario of a thought experiment (COMP), an understanding of what information in the vignette is relevant and which information is irrelevant to making a case judgement (R/IRR), and an acceptance of a broad notion of possibility (POSS). Our analysis of the entire data set indicates that COMP and R/IRR are predictive of the convergence of case judgements on the textbook consensus in both our STRICT and LAX count, and that POSS is predictive of case judgements converging on the textbook consensus in the STRICT count only. This strongly suggests that the skills we identified for the most part constitute skills that are essential for informed and informative case judgements.

We found that philosophers are more likely to converge on the textbook consensus about these judgements than non-philosophers. We also found that philosophers are better than non-philosophes in in R/IRR, POSS, but not COMP. This suggests that philosophers are better than the folk in making case judgements by virtue of their better performance on R/IRR and POSS. They are not better because they comprehend the vignettes more accurately. This is good news, as COMP would seem to be a minimal requirement for delivering reliable case judgements.

Overall, then, philosophers are indeed significantly better than the folk in making case judgements *for the right reasons*: their case judgements are more informed and informative than those of the laypeople.[7] They may therefore rightly be said to be experts

[7] In our study we operationalized only whether case judgements by philosophers and non-philosophers are informed. Trivially, however, if case judgements are informed, they can be used for testing philosophical theories.

in making case judgements. Our results may also be said to lend support to the expertise defense, even though more work is needed for a full vindication of it. In particular, proponents of the expertise defense must not only show that philosophers are experts in case judgements, but also that they are *not as much* subject to the influence of erroneous factors as the folk are.[8] Although we didn't directly test the influence of such factors in our study, our results lend credibility to the expectation that, by virtue of philosophers' better performance on our skill tasks, philosophers should overall be more reliable in their case judgements. In particular, if Machery is right that the entanglement of the target and the superficial content of cases is one of the causes of the unreliability of case judgements (because the superficial content is confusing to subjects), then this expectation is all the more justified: our results show that philosophers are much better than the folk in discerning the target from the superficial content.

We once more hasten to qualify that philosophers may very well be as much subject to erroneous effects in certain classes of thought experiments as the folk. In particular, we do accept that there is now substantial evidence that philosophers are subject to order effects in variants of trolley cases, for example. Still, there is not much evidence that philosophers are subject to such effects outside the moral realm. It is here where our study is most relevant for the expertise defense.

[8] It is important to keep in mind that proponents of the expertise defense have not claimed that philosophical experts are *entirely* immune to the influence of extraneous effects, only that experts are *less* subject to such effects than the folk (Williamson 2011).

Appendix 1

- John is a security guard at a warehouse. One night, the alarm bell rings. John therefore comes to believe that someone has just broken in. Unbeknownst to John, the triggering of the alarm was in fact just the result of some electronic malfunction because the technician messed up. However, as it turns out, there was actually a thief who managed to break into the warehouse at the same time when the alarm rang.
    - JUDGE: John knows that someone has broken into the warehouse. [incorrect]
    - IRR: The technician messed up. / R: The triggering of the alarm was in fact the result of some electronic malfunction.
    - COMP: John's belief that there was a break-in is correct. [incorrect]
    - POSS: The situation described by the scenario is possible. [correct]

- Imagine a woman named Zoe who doesn't know Arabic. Now imagine Zoe sitting in front of a computer screen that displays Arabic symbols. Zoe is given a big book of instructions. It does not contain any grammar rules of Arabic, but only instructions of the form: "if you see a string of symbols *S*, then write this string of symbols *S**". For every symbol she sees on the screen, Zoe's job is to hit the appropriate Arabic letter on her keyboard. What Zoe does not know is that her computer is linked to an internet dating website and that that the letters she types are used in a communication with a native Arabic speaker in Bagdad. The instructions in the big book are so well-devised that the Arabic speaker in Bagdad does not realize Zoe does not know any Arabic.
    - JUDGE: By following the instructions from the book, Zoe understands Arabic. [incorrect]
    - IRR: Zoe is a woman. / R: Zoe doesn't know Arabic.
    - COMP: The Arabic speaker can tell that Zoe does not know Arabic. [incorrect]
    - POSS: The situation described by the scenario is possible. [correct]

- Imagine some scientists wanted to test the psychology of somebody trapped in a room with nothing to taste. Imagine that a man called Eric was born, grew up in that room, and was fed only by intravenous perfusions. Poor Eric never got to taste anything in his entire life. Because Eric has seen the way people react to good food on TV and how they cringe when biting into lemons, Eric becomes intrigued. He studies all the science there is to know about taste. In fact, he becomes the world's leading expert on the matter and knows exactly what goes on inside the brain when one bites into lemons, for example. When Eric is 36 years old, the scientists start to pity Eric and decide to release him from his room. The first thing he asks for is a lemon. He bites into it and for the first time feels how a lemon tastes.
    - JUDGE: Eric learns something new when he bites into a lemon. [correct]
    - IRR: The scientists wanted to study Eric's psychology. / R: Eric never had the experience of tasting anything in his entire life.
    - COMP: Eric knows all the science there is to know about taste. [correct]
    - POSS: The situation described by the scenario is possible. [correct]

- Imagine that the director of a famous Picasso museum wants to test the expertise of the visitors to her museum. To this end, she orders forged copies of the museum's Picasso paintings from the world's best forger. The forger is so good that special instruments are required for detecting the forgeries. The director decides to replace all except one of his thousand Picasso paintings with the forgeries. Jerry is one of the visitors. After he enters the exhibition hall, by a curious coincidence, happens to stop in front of the only real Picasso painting in the entire museum. He then thinks to himself "The painting in front of me is by Picasso".
    - JUDGE: Jerry knows that the painting in front of him is by Picasso. [incorrect]
    - IRR: The director wants to test the expertise of the visitors to her museum. / R: The forger is so good that special instruments are required for detecting the forgeries.
    - COMP: Jerry can tell when the Picasso painting in front of him is forged. [false]
    - POSS: The situation described by the scenario is possible.

- Imagine that a man called Bob hears a catchy song on the radio called “Jingle”. He is told that Jingle was written by Beyoncé, a famous contemporary singer-songwriter. Bob, who does not listen to music much, has never heard of Beyoncé and the only information he has about her is that she wrote Jingle. Now it turns out that the real composer of Jingle is not Beyoncé but by her best friend Ceyoncé, who died a few years ago. Beyoncé just found the score written by Ceyoncé, recorded the song, and took credit for it. Nobody is aware of this.
    - JUDGE: When Bob uses the name “Beyoncé” he is talking about the person who actually wrote the song. [incorrect]
    - IRR: Ceyoncé is Beyoncé’s best friend. / R: The only information Bob has about Beyoncé is that she is the author of the song ‘Jingle’.
    - COMP: The information Bob associates with the name “Beyoncé” is wrong. [correct]
    - POSS: The situation described by the scenario is possible.

- Imagine that way out in our galaxy there is a planet very similar to our planet Earth. Let us call it "Sister Earth". On Sister Earth, a drinkable colourless liquid flows in rivers and lakes and is externally indistinguishable from water on Earth. This liquid, however, is not composed of H2O molecules, but of other molecules that do not exist on Earth. Let us call them XYZ molecules. Now suppose that NASA astronauts managed to visit Sister Earth. The first time they see the watery substance, they shout: "they have water here too".
    - JUDGE: What the astronauts say when they first see the watery substance on Sister-Earth is false. [correct]
    - IRR: Sister Earth is located in our galaxy. / R: The astronauts come from Earth.
    - COMP: The watery substance on Sister Earth, despite not being H2O, looks, smells, and tastes exactly like water on Earth. [correct]
    - POSS: The situation described by the scenario is possible.

Appendix 2

Age

	Min	Max	Median
Philosophers	21	77	35
Non-philosophers	20	72	34

Gender

	Male	Female	Other
Philosophers	176	63	3
Non-philosophers	143	99	0

Philosophy exposure

Answer	**Philosophers**	**Non-Philosophers**
I have a PhD in philosophy	144	0
The highest degree I have in philosophy is a Master's degree and I am enrolled in a PhD programme in philosophy	98	0
The highest degree I have in philosophy is a Master's degree and I am NOT enrolled in a PhD programme in philosophy	0	4
The highest degree I have in philosophy is a Bachelor's degree	0	13
I followed at least one philosophy course at	0	75

university but I do not have a degree in philosophy		
I followed at least one philosophy course in high school but not at university	0	24
I did not follow any philosophy course in high school	0	126

Areas of philosophy

	Philosophers	Non-Philosophers
Epistemology	72	0
Ethics	68	2
History of philosophy	57	0
Metaphysics	54	1
Philosophy of Science	70	0
Philosophy of mind	72	1
Other	61	0

List of other areas (if mentioned more than once the count is in parentheses): Logic (8), Political philosophy (8), Political Theory, Aesthetics (6), Human Condition, Common sense philosophy, Philosophy of Language (23), Philosophy of Law, Emotion, Metaphilosophy (3), Phenomenology (3), Philosophy of art (2), Evolutionary epistemology Philosophy of mathematics (2), Continental philosophy, Philosophy of Medicine, Social philosophy, Critical theory, Indian Philosophy, Legal Philosophy (2)

Education level of non-philosophers

Education Level	Count
1: I did not finish high school	4
2: I finished high school and stopped studying after high school	103

3: I have a Bachelor's degree	97
4: I have a Master's degree	17
5: I have a PhD degree	2

- 2 participants gave contradictory answers and are not included in this table (one answered the first 3 options and the other option 2 and 3)
- 3 participants answered both option 3 and 4 and are here just counted as option 4
- Participant who had either a bachelor's or master's degree in philosophy was not presented with this question (17 in total)

Language proficiency

	Philosophers	Non-Philosophers
Native speaker	102	211
Advanced or mastery	133	24
Intermediate or upper intermediate	7	7

Appendix 3

Negative Binomial Count model of R/IRR, COMP, and POSS with regards to JUDGE **STRICT measure:**

JUDGE = α + R/IRR

	Estimate	Std. Error	z	p
α	0.62	0.07	9.21	<.05
R/IRR	0.15	0.02	7.51	<.05

AIC = 1718.3

JUDGE = α + COMP

	Estimate	Std. Error	z	p
α	0.63	0.07	8.72	<.05
COMP	0.11	0.02	6.78	<.05

AIC = 1727.8

JUDGE = α + POSS

	Estimate	Std. Error	z	p
α	0.88	0.04	20.83	<.05
POSS	0.09	0.02	5.83	<.05

AIC = 1742.2

LAX count

JUDGE = α + R/IRR

	Estimate	Std. Error	z	p
α	1.31	0.07	18.94	<.05
R/IRR	0.04	0.02	2.33	<.05

AIC = 1762.5

JUDGE = α + COMP

	Estimate	Std. Error	z	p
α	1.16	0.1	11.87	<.05
COMP	0.06	0.02	3.15	<.05

AIC = 1757.8

JUDGE = α + POSS

	Estimate	Std. Error	z	p
α	1.37	0.06	22.29	<.05
POSS	0.02	0.01	1.62	ns

AIC = 1765.3

Chi-square tests of all tasks for each scenario

STRICT count

scenario	type	Chi-squared	df	p	% correct phil	% correct non-phil
Gettier	JUDGE	46,04	1	0	47,11	18,18
	IRR	9,33	1	0	13,64	6,2
	R	2,09	1	0,15	27,69	21,9
	COMP	0,53	1	0,47	56,2	52,89
	POSS	48,29	1	0	71,07	39,67
Chinese Room	JUDGE	6,19	1	0,01	84,3	75,21
	IRR	39,79	1	0	42,56	26,03
	R	1,93	1	0,17	35,54	36,36
	COMP	0,68	1	0,41	71,9	75,21
	POSS	15,33	1	0	23,14	9,92
Mary’s Room	JUDGE	0,09	1	0,77	67,77	69,01
	IRR	37,35	1	0	31,4	12,4
	R	2,78	1	0,1	38,84	42,98
	COMP	1,95	1	0,16	42,15	35,95
	POSS	20,06	1	0	18,6	5,37
Fake Barn	JUDGE	16,23	1	0	18,18	6,2
	IRR	13,9	1	0	22,31	11,16
	R	7,7	1	0,01	12,4	20,66
	COMP	29,92	1	0	66,12	41,32
	POSS	18,63	1	0	47,11	28,1
Gödel	JUDGE	0,01	1	0,93	50	49,59
	IRR	40,2	1	0	38,43	15,29
	R	21,29	1	0	4,55	18,18
	COMP	2,18	1	0,14	55,37	61,98
	POSS	37,12	1	0	64,88	37,19
Twin Earth	JUDGE	0,53	1	0,47	43,8	47,11

	IRR	37,9	1	0	37,6	17,36
	R	9,43	1	0	15,29	24,79
	COMP	1,8	1	0,18	71,07	76,45
	POSS	1,5	1	0,22	18,6	14,46

LAX count

scenario	type	Chi-squared	df	p	% correct phil	% correct non-phil
Gettier	JUDGE	56,14	1	0	80,58	47,93
	IRR	8,29	1	0	22,73	14,46
	R	6,96	1	0,01	40,5	31,82
	COMP	4,02	1	0,04	71,49	79,34
	POSS	5,52	1	0,02	92,56	85,95
Chinese Room	JUDGE	6,93	1	0,01	95,87	89,67
	IRR	38,51	1	0	44,21	28,93
	R	1,04	1	0,31	44,21	42,56
	COMP	0,62	1	0,43	84,71	87,19
	POSS	2,68	1	0,1	52,89	45,45
Mary’s Room	JUDGE	0,25	1	0,62	91,32	92,56
	IRR	39,45	1	0	36,78	17,36
	R	0,22	1	0,64	47,93	48,35
	COMP	0,82	1	0,36	73,55	69,83
	POSS	9,7	1	0	43,39	29,75
Fake Barn	JUDGE	30,59	1	0	50	25,62
	IRR	20,97	1	0	33,06	18,6
	R	1,57	1	0,21	28,1	31,82
	COMP	18,13	1	0	82,64	65,7
	POSS	1,38	1	0,24	78,51	73,97
Gödel	JUDGE	0,65	1	0,42	73,14	69,83

	IRR	33,36	1	0	43,39	22,31
	R	37,71	1	0	6,2	26,45
	COMP	0,14	1	0,71	84,71	83,47
	POSS	11,29	1	0	88,43	76,86
Twin Earth	JUDGE	0,09	1	0,76	70,66	71,9
	IRR	37,83	1	0	43,39	23,97
	R	4,03	1	0,04	30,17	35,95
	COMP	10,43	1	0	84,71	93,8
	POSS	0,41	1	0,52	45,45	48,35

t-tests of **STRICT** scores for the different question **types.**

Type	Phil, mean	SD	Non-phil, mean	SD	t	df	p
JUDGE	3,11	1,49	2,65	1,41	3,48	480,46	< .05
R/IRR	3,2	1,29	2,53	1,41	5,45	477,75	< .05
COMP	3,63	1,63	3,44	1,6	1,3	481,87	ns
POSS	2,43	1,79	1,35	1,38	7,48	452,51	< .05

t-tests of **LAX** scores for the different question **types.**

Type	Phil, mean	SD	Non-phil, mean	SD	t	df	p
JUDGE	4,62	1,12	3,98	1,16	6,2	481,47	< .05
R/IRR	4,21	1,12	3,43	1,37	6,85	463,53	< .05
COMP	4,82	1,17	4,79	1,16	0,23	481,95	ns
POSS	4,01	1,53	3,6	1,43	3,04	479,71	< .05

References

Alexander, J. and J.M. Weinberg. 2007. Analytic epistemology and experimental philosophy. *Philosophy Compass*, **2** (1): 56-80.

———. 2014. The "unreliability" of epistemic intuitions. In *Current Controversies in Experimental Philosophy*, Edouard Machery and Elizabeth O'Neill (eds.), New York: Routledge, 128-145.

Bourget, D. and D.J. Chalmers. 2014. What do philosophers believe? *Philosophical Studies*, **170** (3): 465-500.

Chalmers, D.J. 2015. Why Isn't There More Progress in Philosophy? *Philosophy*, **90** (1): 3-31. https://www.cambridge.org/core/article/why-isnt-there-more-progress-in-philosophy1/DD29F4A95066E4D628F097AE5EF53DB3.

Chi, M.T., R. Glaser, and M.J. Farr. 2014. *The nature of expertise*. New York: Psychology Press.

Cole, D. 2015. The Chinese Room Argument. *The Stanford Encyclopedia of Philosophy (Winter 2015 Edition)*, edited by Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2015/entries/chinese-room/

Cova, F., B. Strickland, A.G.F. Abatista, A. Allard, J. Andow, M. Attie, J. Beebe, R. Berniūnas, J. Boudesseul, and M. Colombo. preprint. Estimating the reproducibility of experimental philosophy.

Dennett, D.C. 1993. *Consciousness explained*: Penguin UK.

Devitt, M. 2011. Experimental semantics. *Philosophy and Phenomenological Research*, **82** (2): 418-435.

———. 2012a. Semantic Epistemology. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia*, **27** (2): 229-233.

———. 2012b. Whither experimental semantics? *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia*, **27** (1): 5-36.

Ericsson, K.A. 2006. The influence of experience and deliberate practice on the development of superior expert performance. *The Cambridge handbook of expertise and expert performance*, **38**: 685-705.

Feltz, A. and E.T. Cokely. 2012. The philosophical personality argument. *Philosophical Studies*, **161** (2): 227-246.

Häggqvist, S. 2009. A model for thought experiments. *Canadian Journal of Philosophy*, **39** (1): pp. 55-76.

Hales, S.D. 2006. *Relativism and the Foundations of Philosophy*. Cambridge, MA: MIT Press.

Heil, J. 2013. *Philosophy of mind: A contemporary introduction*: Routledge.

Horvath, J. 2010. How (not) to react to experimental philosophy. *Philosophical Psychology*, **23** (4): 447-480.

Horvath, J. and A. Wiegmann. 2015. Intuitive expertise and intuitions about knowledge. *Philosophical Studies*: 1-26.

Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly*, **32** (April): 127-136.

Lowe, E.J. 2000. *An introduction to the philosophy of mind*. Cambridge: Cambridge University Press.

Ludwig, K. 2007. The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy*, **31** (1): 128-159.

Machery, E. 2012. Expertise and intuitions about reference. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia*, **27** (1): 37-54.

———. 2017. *Philosophy within its proper bounds*: Oxford University Press.

Machery, E., R. Mallon, S. Nichols, and S.P. Stich. 2004. Semantics, cross-cultural style. *Cognition*, **92** (3): B1-B12.

Nado, J. 2014. Why intuition? *Philosophy and Phenomenological Research*, **89** (1): 15-41.

Schulz, E., E.T. Cokely, and A. Feltz. 2011. Persistent bias in expert judgments about free will and moral responsibility: A test of the expertise defense. *Consciousness and Cognition*, **20** (4): 1722-1731.

Schwitzgebel, E. and F. Cushman. 2012. Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, **27** (2): 135-153.

———. 2015. Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition*, **141**: 127-137.

Searle, J.R. 1980. Minds, brains, and programs. *Behavioral and brain sciences*, **3** (3): 417-424.

Thagard, P. 2014. Thought experiments considered harmful. *Perspectives on Science*, **22** (2): 288-305.

Tobia, K., W. Buckwalter, and S. Stich. 2013. Moral intuitions: Are philosophers experts? *Philosophical Psychology*, **26** (5): 629-638.

Van Inwagen, P. 2004. Freedom to break the laws. *Midwest Studies in Philosophy*, **28** (1): 334-350.

Weinberg, J.M., C. Gonnerman, C. Buckner, and J. Alexander. 2010. Are philosophers expert intuiters? *Philosophical Psychology*, **23** (3): 331-355.

Wiegmann, A., J. Horvath, and K. Meyer. forthcoming. Intuitive Expertise and Irrelevant Options. In *Oxford Studies in Experimental Philosophy*, Tania Lombrozo, Joshua Knobe and Shaun Nichols (eds.).

Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.

———. 2011. Philosophical expertise and the burden of proof. *Metaphilosophy*, **42** (3): 215-229.