Philosophical Criticisms of Experimental Philosophy

(version of January 2015; to appear in J. Sytsma and W. Buckwalter (eds.), *A Companion to Experimental Philosophy*, Wiley Blackwell.

Timothy Williamson

Abstract:

The philosophical relevance of experimental psychology is hard to dispute. Much more controversial is the so-called negative program's critique of armchair philosophical methodology, in particular the reliance on 'intuitions' about thought experiments. This chapter responds to that critique. It argues that, since the negative program has been forced to extend the category of intuition to ordinary judgments about real-life cases, the critique is in immediate danger of generating into global scepticism, because all human judgments turn out to depend on intuitions. Recently, some proponents of the negative program have tried to demarcate the target of their critique more narrowly. However, their attempts are still far too indiscriminate, and over-generate scepticism. Nevertheless, once experimental philosophy has refined its own methodology, it may contribute to the refinement of the methodology of mainstream philosophy, by filtering out the effects of cognitive bias, although it offers no prospect of doing without judgments on real or imaginary cases.

Keywords:

Experimental philosophy, metaphilosophy, philosophical methodology, thought experiment, intuition, epistemology, scepticism

1. Introduction

The phrase ‘experimental philosophy’ can mean many things. In a broad sense, it covers any experimental inquiry with a philosophical purpose (Rose and Danks 201X argue for a similarly broad understanding of ‘experimental philosophy’). On that reading, few philosophers today object to experimental philosophy as such. For example, it is generally agreed that the philosophy of perception has much to learn from experiments on the psychology of perception. Although the experiments tend to have been conducted by psychologists for psychological rather than philosophical purposes, in principle philosophers of perception themselves could initiate and even conduct similar experiments for philosophical purposes — although in practice the results will be better if they do so in collaboration with experimental psychologists, who have more of the required know-how in designing, conducting, and interpreting experiments. Analogous considerations apply to the philosophy of space and time and experiments in physics. A few diehard Wittgensteinians may still claim that no outcome of scientific experimentation is of special relevance to philosophy, whose role they confine to dissolving conceptual confusions. This chapter assumes that philosophy is a theoretical discipline with more constructive ambitions than that.

In a narrower sense, ‘experimental philosophy’ refers to a more specific kind of philosophically motivated experimental inquiry, in which verdicts on hypothetical cases relevant to some philosophical question are elicited from significant numbers of subjects, sometimes under controlled conditions, and hypotheses are tested about the underlying

patterns. Again, there is no reason in principle why philosophy cannot learn from the results of such activities, though their bearing on the original philosophical questions needs to be clarified. But within experimental philosophy in the narrower sense, there is a minority movement, sometimes known as the 'negative program', which has attracted attention disproportionate to its size, because its proponents' claims seem to have radical implications for philosophical methodology. The negative program offers a naturalistic critique of the non-experimental ('armchair') methods of much recent analytic philosophy, and in particular of its reliance on thought experiments (for these purposes, thought experiments do not count as experiments themselves). The well-known paper by Weinberg, Nichols, and Stich (2001) may conveniently be taken as the opening broadside of the negative program, at least in its contemporary form. The results of some of its experiments are interpreted as showing that the crucial verdicts in thought experiments on which philosophers have relied are sensitive to factors presumably irrelevant to their truth, such as the ethnicity or gender of the experimental subjects, or the order or environment in which they are presented with the thought experiments. Although most experimental philosophy even in the narrow sense is independent of that critique, this chapter focusses on the negative program, and criticisms of it. Nor does it concern all such criticisms. Various objections have been raised to the specific design, execution, interpretation, and repeatability of specific experiments on which proponents of the negative program have relied. This chapter does not discuss such objections (but see CHAPTER XX (ARMCHAIR FRIENDLY EXPERIMENTS)). Rather, it concentrates on broader theoretical challenges to the negative program that arise even if the specific experiments at issue are well designed, well executed, well interpreted, and repeatable.

2. ‘Philosophical Intuitions’

Many proponents and many opponents of philosophical thought experiments describe them as eliciting ‘philosophical intuitions’, corresponding to the crucial verdicts. For example, it is said to be a philosophical intuition that, in the hypothetical scenario, the subject ought to divert the trolley to save five lives at the expense of one, or does not know that it is 3 p.m. by looking at a stopped clock that happens to be showing the right time. Thus many proponents of the negative program define the overall target of their methodological critique as reliance on philosophical intuitions, or on intuitions more generally (see e.g. Alexander and Weinberg, 2007, 63). Against them, many other philosophers defend reliance on philosophical intuitions, or on intuitions more generally (Sosa 2007). Still others deny that philosophical thought experiments involve reliance on such intuitions (Deutsch 2009, Cappelen 2012).

The phrase ‘philosophical intuition’ is obviously technical jargon, in need of explanation. Surprisingly, both proponents and opponents of the negative program tend to use the phrase as though it were self-explanatory. Alternatively, they give it a perfunctory vague gloss such as ‘*what we would say* or *how things seem to us*’ (Alexander 2012, 1). At first sight, this does not look like much of a problem, since it seems clear enough from examples what is meant. We can recognize a philosopher’s thought experiment when we see one, and the crucial verdict in it is the one the philosopher subsequently invokes. Of course, examples do not reveal the underlying psychological nature of philosophical

intuitions, but we need not know that nature in order to recognize when they are being relied on.

We can start to appreciate the inadequacy of that attitude by considering real life analogues of thought experiments. In epistemology, I have sometimes played tricks on audiences to create actual Gettier cases (Williamson 2007, 192). Instead of judging that in the hypothetical scenario the subject has justified true belief without knowledge of the given fact, audience members judged (after I revealed the trick) that they themselves had had justified true belief without knowledge. Instead of judging that the man you *imagine* relying on a stopped clock would not know that it is 3 o'clock, you can judge that the man you *observe* relying on a stopped clock does not know that it is 3 o'clock. Indeed, you can receive the description of the imaginary case in the very same words as a report of a real life case, and judge whether he knows on that basis. For epistemological purposes, such actual cases do just as well as hypothetical ones in showing justified true belief to be insufficient for knowledge.[1] If epistemologists rely on actual cases rather than hypothetical ones, are they still relying on philosophical intuitions? If the negative program's answer is 'No', its critique of reliance on philosophical intuitions will be quite easy to get round in some key debates: just bring about a real life analogue of the contested thought experiments. Of course, that will often be a laborious business, and in moral philosophy an unethical one, since lives will be lost in the non-fictional analogues of trolley cases. Nor is such an alternative available for the more science fictional cases. Nevertheless, for some of the thought experiments which negative programmers have expended most effort resisting, their resistance would have been futile.

Understandably, negative programmers have preferred to rule that using real life cases instead of the corresponding philosophical thought experiments still counts as relying on philosophical intuitions. That ruling is not *ad hoc*. It is very plausible that the cognitive processes underlying the crucial verdicts on the imagined hypothetical cases have much in common with the cognitive processes underlying the crucial verdicts on the corresponding experienced real life cases (Williamson 2007, 179-207). Thus it is natural for negative programmers to extend suspicion of the cognitive processing of imaginary cases to suspicion of the cognitive processing of corresponding real-life cases, since one might expect biases in the former to be inherited from similar biases in the latter. However, this extension has a price. Our fairly straightforward ability to discriminate situations where thought experiments are being performed from other situations no longer constitutes an ability to discriminate situations where *philosophical intuitions* are being used from other situations, since many situations where philosophical intuitions about real life cases are being used are situations where no thought experiment is being performed. For virtually any judgment one makes on an actual case, there is a corresponding judgment to be made on an analogous hypothetical case, and using that hypothetical case for a thought experiment may suit the dialectical purposes of some philosopher, since some other philosopher may have inadvertently proposed a theory to which it is a counterexample. The obvious danger is that the category of philosophical intuitions will be stretched so wide, encompassing virtually anything one says about actual cases, that the negative programmers' critique of reliance on philosophical intuitions will become a global scepticism, at odds with their conception of their general enterprise as a positive contribution to naturalistic inquiry.

Can negative programmers reply that what counts as a 'philosophical intuition' is itself a matter for further experimental inquiry to determine, by uncovering underlying

similarities? The trouble with such a reply is that negative programmers take their critique in its current state *already* to have present practical implications for philosophical methodology. They face the challenge of articulating those implications without assuming that we are already in a position to recognize a philosophical intuition when we see one. As already explained, the methodological ambitions of the negative program require us to reform our practices with respect to real life cases as well as fictional ones, but they leave it quite unclear how much they intend that category to include.

In the heady early days of the negative program, a commonly drawn moral was that philosophers should *stop relying* on philosophical intuitions, at least until substantial experimental evidence was produced of their reliability. But how can philosophers act now on that moral if they have no idea how far the category of philosophical intuitions extends? If negative programmers are banning some parts of current philosophical practice, they had better make it clear enough for present working purposes which parts they are banning. Thus, if they define those parts as the ones that involve reliance on 'philosophical intuitions', they had better make it clear enough for present working purposes which situations involve reliance on philosophical intuitions. Mere appeal to the results of future experimental inquiry is not enough for present working purposes.

Since those early days, negative programmers have become more cautious, in response to both philosophical criticisms and difficulties in reproducing experimental results. As noted earlier, there is an increasing realization that the category of 'philosophical intuitions' may be so broad that general scepticism about them can easily lead to hopeless global scepticism. A view something like the following is now widespread: The overall reliability of philosophical intuitions may well be quite high: non-accidentally, a reasonable

proportion of them are true. However, such moderate global reliability is consistent with both extreme local unreliability here and there, and less extreme but more global unreliability more widely, resulting from bias, distortion, and sensitivity to irrelevant factors. In the background of this picture may be an evolutionary line of thought: for central, common cases our practices of applying a concept have enough causal repercussions for a propensity to serious error to have a significant cost in fitness, but in rare or marginal cases that is not so.

One might try stating the proposed methodological moral of the negative program in a more circumscribed way: we should not rely on a *specific* philosophical intuition until we have experimental evidence that *it* is widely shared. However, the difficulty remains: how can we act on that advice unless we can recognize a philosophical intuition when we see one?

The difficulty depends on the presumption that the methodological moral is not being generalized *beyond* the category of philosophical intuitions. If mad-dog naturalists make such a generalization, and insist that we should not rely on any judgment at all until we have experimental evidence that it is widely shared, it may not matter for their purposes whether the judgment counts as a philosophical intuition. But the more general moral is hopeless, because it generates an infinite regress: the experimental evidence takes the form of a report of the experiment, that report consists of the authors' judgments, on which we are told not to rely until we have experimental evidence that *they* are widely shared, and so on. Negative programmers do not endorse such mad-dog generalized morals. Their methodological moral is specific to philosophical intuitions, which is why its application

depends on our ability to distinguish in practice between philosophical intuitions and other judgments.

Not all negative programmers insist that we must wait until we have experimental evidence that a philosophical intuition is widely shared before we rely on it. A more moderate moral is that we may rely on a philosophical intuition even in the absence of experimental evidence that it is widely shared, as long as no one rejects the intuition — but once someone has rejected it, we must suspend judgment on it until we get such experimental evidence. But the workability even of that more moderate moral depends on our ability to distinguish in practice between philosophical intuitions and other judgments, unless the moral is generalized to those other judgments. Once again, the generalized moral is hopelessly immoderate. It implies that we must suspend any judgment that someone has rejected until we have experimental evidence that it is widely shared. That principle would make it all too easy for a troublemaker to bring any inquiry he or she disliked to a grinding halt, simply by rejecting a key judgment on which its practitioners relied, then rejecting a key judgment in the report of the experimental evidence that the former judgment was widely shared, and so on. In particular, such a malicious critic could soon stop the negative programme in its tracks.

The methodological moral can be watered down still further, so that more than one lone troublemaker is required to trigger the obligation to suspend judgment until experimental evidence is obtained. But numbers are not the issue: naturalists cannot accept any generalized methodological moral that would enable large teams of postmodernists or religious fundamentalists to bring natural scientific inquiry to a standstill just by rejecting key judgments whenever it suited them, in order to trigger a potentially infinite regress of

experimental demands. Thus the point remains: the intended methodological moral of the negative program mandates some sort of special treatment for a category of 'philosophical intuitions', so its present workability depends on our present ability in practice to determine when we are faced with a member of that category. Negative programmers are treating disagreement in philosophical intuitions differently from disagreement in other judgments. They cannot simply sidestep the demand for a workable demarcation of the category. What differentiates philosophical intuitions from the rest?

There is no promise in the idea of distinguishing *philosophical* intuitions by something distinctively philosophical in their *content*. The only candidate in the content of the supposed philosophical intuition 'He doesn't know that it's 3 p.m.' is the reference to knowledge, a philosophically interesting relation. But if the use of the ordinary term 'know' for a philosophically interesting relation suffices to make 'He doesn't know that it's 3 p.m.' a philosophical intuition, then the discourse of experimental philosophers themselves is packed with philosophical intuitions, since they often apply ordinary terms such as 'learn' (acquire knowledge) and 'evidence' for philosophically interesting relations to specific cases. The problem of over-generation remains.

Intuitive judgments are often contrasted with *reflective* judgments (see e.g. Nagel 2012: 497-503, drawing on Mercier and Sperber 2009). The difference is not that reflective judgments are based on evidence, for so are many intuitive judgments. Thus the 'philosophical intuition' in a real life Gettier case 'He doesn't know that it's 3 p.m.' depends on evidence such as that the clock he looked at has stopped, that he is wearing no watch, and so on. In the corresponding thought experiment, 'He doesn't know that it's 3 p.m.' presumably relies on hypothetical evidence in a similar way, and when one steps back

outside the imaginative exercise to judge 'In the story, the man doesn't know that it's 3 p.m.', that does not undo the original use of evidence; it simply involves a further step of conditionalization, marked by the introduction of the operator 'in the story'. The difference is rather that reflective judgments are reached through something like consciously controlled reasoning, in a series of steps, whereas intuitive judgments are not. For instance, if one reasons to oneself 'No one who relies on a stopped clock knows the time; he is relying on a stopped clock; therefore he does not know that it is 3 o'clock', the concluding judgment is reflective rather than intuitive. Consciously controlled reasoning has distinctive psychological features: unlike intuitive judgment, it is slow, it makes heavy demands on working memory, and it can only integrate very limited amounts of information.

By the proposed standard, the judgment about the thought experiment 'In the story, the man doesn't know that it is 3 o'clock' may count as *less* intuitive than the judgment about the real life case 'He doesn't know that it is 3 o'clock', since the former but not the latter involves the extra step of conditionalization noted above, marked by 'in the story', which may well be a piece of consciously controlled reasoning. However, we can allow that there is a spectrum from intuitive judgments through increasingly reflective ones, and that here we are still close to the intuitive end. But grading intuitiveness does not mean that the negative program can confine itself to judgments that are not highly reflective. For example, having judged 'He doesn't know that it is 3 o'clock', through a series of steps of consciously controlled reasoning one can conclude 'A son of a child of a child of that man's great-grandmother in the maternal line has a justified true belief that it is 3 o'clock without knowing that it is 3 o'clock', which counts as a highly reflective judgment by the proposed standard. It does just as well as the original intuitive judgment for arguing against the justified true belief account of knowledge.

Clearly, the negative program needs to extend to reflective judgments derived from intuitive judgments. But what reflective judgments are *not* derived from intuitive judgments? If a reflective judgment results from several steps, what about the *first* judgment in the series? Suppose that one reflectively concludes 'Socrates is mortal' by syllogistic reasoning from 'All men are mortal' and 'Socrates is a man'. One's judgment 'Socrates is a man' may well be intuitive; if one consciously recognizes the valid pattern of the reasoning, the judgment in which one does so may also count as intuitive. If those judgments are not intuitive, others earlier in the process will be. As the distinction has been drawn, all reflective judgments rely on intuitive judgments. If intuitive judgments are the outputs of system 1 and reflective judgments of system 2, the point is that all system 2 thinking involves system 1 thinking. Thus scepticism about intuitive judgments generalizes to scepticism about *all* judgments. It is an illusion that reliance on intuitive judgments, characterized along anything like the lines sketched above, constitutes a distinctive method of armchair philosophy. In that sense of 'intuitive', all human thinking relies on intuitive judgments.

Both opponents and proponents of a postulated distinctively philosophical method of 'reliance on intuitive judgments' need to demarcate 'intuitive judgment' much more narrowly. Another sign of this is that ordinary perceptual judgments come out as intuitive rather than perceptual, but they are far from the only non-reflective judgments that are not supposed to be at issue. Even mathematical reasoning ultimately relies on non-reflective pattern recognition. But it is quite unclear how this required narrower type of 'intuitive judgment' is supposed to be demarcated.

Unfortunately, the terms ‘intuition’ and ‘intuitive’ continue to be used by all sides in debates on philosophical methodology without remotely adequate clarification. This is a significant obstacle to progress. A more hopeful sign is that some negative programmers have seen the need for a much more nuanced and qualified characterization of the target of their methodological critique, one that puts little or no weight on the category of philosophical intuition. Such a redefinition of the terms of debate should facilitate progress. The next section discusses the redefined debate.

3. Proper Domains for the Application of Concepts

For definiteness, I will concentrate on a paper by Edouard Machery (2011) that argues for the combination of moderate global reliability with local unreliability in the setting of the negative program. To his credit, Machery avoids the term ‘intuition’ altogether, so the concerns of section 2 do not arise directly for him.

Machery is sympathetic to what he calls ‘the Ordinary Judgment Proposal’, that ‘the judgments elicited by thought experiments are underwritten by the psychological capacities that also underlie the judgments we make about everyday situations’ (2011, 194). What he calls ‘the Parity Defense of Thought Experiments’ argues from the Ordinary Judgment Proposal to the conclusion that one cannot challenge the ‘reliability and thus trustworthiness’ of the judgments elicited by thought experiments ‘without also challenging the reliability and thus trustworthiness of all our judgments—a price too high to pay for

even the most ardent critics of thought experiments' (2011, 196). Machery attacks the Parity Defense, and indeed argues that the Ordinary Judgment Proposal has sceptical implications for philosophical thought experiments (2011, 197).

According to Machery, 'the main criticism of the Parity Defense' is that we have reason to believe that philosophical thought experiments involve the application of concepts in situations outside the proper domain of the psychological capacities underlying our application of those concepts, where the proper domain of a psychological capacity is defined to comprise the circumstances in which it is reliable (2011, 201). Machery is obviously right that the Ordinary Judgment Proposal does not *entail* the Parity Defense. It is logically consistent to hold that the psychological capacities underlying our application of a given concept are reliable in everyday situations but unreliable in philosophical thought experiments. The question is whether we have any reason to believe that combination of claims, and in particular whether the Ordinary Judgment Proposal gives us any reason to believe it.

The mere atypicality of the circumstances does not give us good reason to believe that we are outside the proper domain of the relevant concept. Atypicality does not imply unreliability. For example, some people have exceptionally good memories; they are good to a rare, atypical degree. That does not give us reason to believe that we are outside the proper domain of the concept of remembering when we apply it to them. Although atypicality may tend to increase the chance of unreliability, it does not in general do so enough to warrant agnosticism. After all, situations of danger tend to be atypical in various ways; we are in trouble if our cognitive systems fail whenever we need them most.

Machery himself is sometimes quite liberal about proper domains. 'At an abstract level', he says, the situations described in science fiction novels 'are clearly very similar to everyday situations, and we thus have reason to believe that they belong to the proper domains of the relevant psychological capacities' underlying our judgments about those science fictional situations (2011, 202 n11). In Machery's view, the most important characteristic of philosophical thought experiments in giving us reason to believe that they fall outside the proper domains of the relevant concepts is that they 'typically pull apart the features that go together in everyday life' (2011, 203). As he points out, if the imagined cases have this characteristic, then their real life counterparts will share it.

Machery's first example is that in a standard thought experiment from moral philosophy (pushing a fat man off a footbridge to save five other people), 'using physical violence and doing more harm than good are pulled apart', whereas using physical violence and doing more harm than good supposedly go together in everyday life. Thus, his argument goes, we have reason to believe that the psychological capacities underlying our application of moral concepts are unreliable in such cases, and therefore to be sceptical about our initial moral judgment. But consider a woman who fights off her would-be rapist, kicking him in the groin and having him arrested. We judge that her action was morally permissible, indeed right. But this too is a case of using physical violence without doing more harm than good, and therefore pulls apart the features that go together in everyday life. According to Machery's argument, therefore, we have reason to believe that the psychological capacities underlying our application of moral concepts are unreliable in this case too, and therefore to be sceptical about our initial judgment that the woman's action was morally permissible. Surely this scepticism is unwarranted, and potentially pernicious. More generally, although professors at top universities may rarely encounter at first-hand

situations in which physical violence is the only effective form of self-defence or defence of innocent people, such situations have been quite common in human experience. Thus Machery's argument as he states it severely over-generates scepticism about moral judgment. No doubt it is rare to be able to save many people by killing one, but to characterize the supposedly problematic feature of the case so narrowly would smack of special pleading.

The treatment of epistemologists' thought experiments is similar. According to Machery (2011, 204):

> When people fail to know something, their beliefs are typically false, unjustified, and the products of unreliable methods. When people know something, their beliefs are typically true, justified, and the product of reliable methods. By contrast, Gettier cases sever truth and justification from the reliability of the methods of belief formation since they describe situations where truth comes about by luck. [Footnote: Here the method is not the tendency to endorse one's perceptual experience (which is a reliable method) but the use of a broken clock.] Thus, one has a reason to believe that the situations described by Gettier cases are beyond the proper domain of our everyday capacity to ascribe knowledge.

Here Machery seems to assume that we have a reason to believe that any situation where the three features of truth, justification, and reliability of the methods of belief formation fail to go together is beyond the proper domain of our everyday capacity to ascribe knowledge (or its absence). Therefore we should be sceptical about our initial judgment that the protagonist of the Gettier case lacks knowledge. Now consider a man who irrationally forms beliefs simply on his guru's authority. The guru makes assertions at random; a few of

them are true, so the follower forms some true beliefs. Those cases sever truth from justification and the reliability of the methods of belief formation. Therefore, by the principle on which Machery seems to be relying, we have a reason to believe that the situation of the follower's true beliefs is beyond the proper domain of our everyday capacity to ascribe knowledge or its absence. Therefore we should be sceptical about any initial judgment we may have made that the follower lacks knowledge. Again, this scepticism is surely unwarranted. Thus Machery's argument severely over-generates scepticism about epistemological judgment.[2]

Machery's takes the same line about the sort of thought experiment that Kripke (1980) uses to refute descriptivist theories of reference for proper names (2011, 204):

> Situations involving proper names associated with a single description that happens to be false of the original bearer of the name are probably beyond the proper domain of our capacity to identify the reference of proper names since in everyday circumstances many of the numerous descriptions associated with a proper name tend to be true of the original bearer of the name.

But it is just false that in everyday circumstances numerous descriptions are always associated with a proper name. Think of the proper names we picked up when half-attending to lessons in schools, the conversations of others, the television, or the internet, and subsequently forgot the source (as often happens to me). Kripke's examples are of an utterly familiar type, slightly schematised only to make the point clearer. For instance, someone uninterested in sport may associate only the description 'professional soccer player' with the name 'Toby Flood' and falsely believe 'Toby Flood is a professional soccer player'; in fact, 'Toby Flood' refers to a professional rugby union player (meta-linguistic

descriptions like ‘the person called “Toby Flood”’ need special discussion, which Kripke (1980) gives them). Such cases occur frequently in everyday circumstances. Here Machery’s argument over-generates scepticism about semantic judgment.

Although the psychological capacities underlying our application of ordinary concepts are doubtless unreliable in some circumstances, Machery’s diagnostics for falling outside their proper domain are far too weak to provide good evidence unreliability. They severely underestimate the range of variation amongst the cases with which we need to deal reliably in everyday life. Animals need minds in order to deal flexibly and appropriately with the somewhat complex, novel situations they not infrequently find themselves in. A high proportion of ordinary cases are complex enough to fit Machery’s diagnostics. For instance, he gives this example of a reliable everyday judgment about knowledge: ‘judging by her answer to the test, one of my undergraduate students does not know what the DN account of explanation is’ (2011, 195-6). By the loose standards Machery applies in assessing philosophical thought experiments, the features of lacking elementary knowledge in an academic field and of never having taken a course on it ‘typically’ go together in everyday life, but they pull apart in this case, so we have reason to believe that the psychological capacities underlying his judgment that his student does not know what the DN account of explanation is are being applied outside their proper domain, and we should be sceptical of his judgment. Once again, his style of argument severely over-generates scepticism. Far more exacting criteria would be needed to provide serious reason to expect unreliability in a given case. Machery does not offer such criteria. Since the Ordinary Judgment Proposal is in no way committed to his easy-going criteria that over-generate scepticism about judgment, Machery’s claim that it implies scepticism about philosophical thought experiments is unfounded.

Like Machery, Joshua Alexander and Jonathan Weinberg (2014) defend a qualified version of the negative program. Unlike him, they still make frequent use of the unclarified term 'intuition'. Concerned to avoid global scepticism, they envisage intuitions as moderately reliable in general but subject to various potential sources of error over which, they claim, only experimental methods will give us control.

Alexander and Weinberg propose some specific features of thought experiments that we might take as danger signals of a potential error source. For instance, they suggest as such a danger signal that the reader of many epistemological thought experiments is supplied with more information about their protagonists' mental states than is typically available in everyday life. That is true; philosophers supply such information in hopes of making their thought experiments as watertight as possible. However, Alexander and Weinberg give no evidence that supplying less information would make a significant difference to the outcome. For example, the man who truly believes that it is 3 p.m. by looking at a stopped clock can be described from the perspective of an external observer watching the man. That does not reverse the verdict that he does not know that it is 3 p.m.

Alexander and Weinberg suggest that the 'subtle' or 'unusual and marginal sorts of cases that are popular with epistemologists' are prime candidates for local unreliability, although they also allow that some sources of bias may be present in more ordinary cases too, and that our 'intuitions' may sometimes withstand experimental tests even in extraordinary cases. They do not expand on what it takes for a case to be 'subtle' or 'marginal'. As for 'unusual', their use of the term is vulnerable to the problem of generality. Any case whatsoever falls under many descriptions, some more specific than others, and so belongs to many sorts. The narrowest sorts to which it belongs will be highly unusual ones;

however ordinary the case, a sufficiently fine-grained description of it will apply to few or no actual cases. At the other extreme, the broadest sorts to which the case belongs will be very usual ones; however extraordinary the case, a sufficiently coarse-grained description of it will apply to many actual cases. In Machery's phraseology, at an abstract enough level the situations described in epistemological thought experiments are clearly very similar to everyday situations, just as the situations described in science fiction novels are. At a less abstract level, in practice every application of a concept is made in a situation different in some respects from all previous situations. The action is in the sorting of cases in the first place, which Alexander and Weinberg fail to discuss. The sorts need to be individuated in such a way that the differences between them may reasonably be expected to correlate with differences in the reliability of the relevant psychological capacities. Without such a principle of individuation, the emphasis on the rarity of the sorts of cases to which epistemologists appeal is just the kind of generic sceptical move that will discredit the experimental philosophers' critique.[3]

One consequence of this failure to provide useful danger signals of unreliability is that it remains unclear what methodological moral philosophers are supposed to draw from the negative critique. 'Avoid unusual, marginal, or subtle cases!' is not very helpful advice. After all, compared to everyday life, a carefully controlled experiment looks like an unusual, marginal, and subtle sort of case, but presumably we are allowed to apply ordinary epistemological concepts such as 'evidence' and 'learning' to it. One challenge to the negative program is to provide a much clearer, more workable and less generic specification of what are supposed to be the serious danger signals.

4. Further Questions about the Parity Defense

Machery (2011) raises several other interesting issues about the Parity Defense of Thought Experiments, which this section will discuss.

Machery reasonably points out that if psychological capacities underlying the application of a concept are unreliable in everyday life, the Ordinary Judgment Proposal suggests that they will be unreliable in thought experiments too. We cannot normally expect imagination to do better than observation. So far so good. Moreover, he argues, 'everyday causal judgments in the social domain are biased, and they are unlikely to be reliable' (a sweeping generalization for which he provides minimal evidence). He concludes that 'causal judgments elicited by thought experiments provide no evidence for the premises of philosophical arguments when the judgments bear on whether an agent caused an outcome' (2011, 200).[4]

Once again, Machery's argument severely over-generates scepticism. Consider this thought experiment:

> Life has not advanced beyond stone-age technology. A community has been living on an island for many years without communicating with the rest of the world. A woman there utters a word. A second later, a man ten thousand miles away utters another word. Did her utterance cause his utterance?

Presumably, we judge that the answer is 'No'. That judgment 'bears on whether an agent caused an outcome'. Therefore, given Machery's conclusion, that judgment should not be

relied on in philosophical argument. This seems rather extreme. To vary the example, consider Machery's own case above of a *reliable* everyday judgment: judging by her answer to the test (he uses the female pronoun), he judges that one of his undergraduate students does not know what the DN account of explanation is. That judgment is in the social domain, and it depends on the causal judgment that her bad answer was caused by her ignorance rather than by her determination to get a bad grade in order to win a bet. Should we therefore reclassify the judgment as unreliable? Presumably not. What all this really shows is again that one must take much greater care to avoid more or less generic scepticism about judgment.

What Machery calls his least important criticism of the Parity Defense is that some philosophical thought experiments have no counterparts in everyday life because they involve matters that lay people do not consider (2011, 197-8). That may be so. For example, some thought experiments about reference may involve a more theoretically constrained reading of 'reference' than is employed in everyday life — and they may be none the worse for that, if the theoretically constrained reading is clear.

Machery's own example of the point is Burge's arthritis thought experiment (Burge 1979). We are to imagine two situations, in which the medically untrained protagonist (Oscar) is in all the same internal physical states and sincerely says 'I have arthritis in my thigh'. The underlying difference between the two situations is in how the rest of Oscar's speech community uses the word 'arthritis'. In situation S1, they apply it as in the actual world only to arthritis, an ailment of the joints but not of the thighs. In situation S2, they apply it much more broadly, to both ailments of the joints and ailments of the thighs. Burge argues that Oscar's beliefs differ in content between the two situations — in S1 but not in S2

Oscar believes that he has arthritis — and therefore that the contents of propositional attitudes do not always supervene on internal physical states, but may depend on the external social environment. Machery complains that since lay people do not consider the individuation of the content of propositional attitudes, the psychological capacities used in everyday life do not support Burge's thought experiments. We can certainly grant Machery that asking theoretical questions about the individuation of content is no part of everyday life. But that is far less damaging to Burge's thought experiment than Machery assumes. Note first that Oscar does not have arthritis in his thigh in either S1 or S2, since it is a medical fact that one cannot have arthritis in one's thigh. After all, Oscar does not have arthritis in his thigh in the straightforward situation S1, and he is in exactly the same medical state in S2 as in S1, so he does not have arthritis in his thigh in S2. Note second that in S1 Oscar believes that he has arthritis in his thigh. This is an everyday propositional attitude ascription, reporting the sort of ordinary medical error to which non-experts are prone. Machery himself describes Oscar in S1 as 'convinced that he has arthritis in his thigh' (2011, 197). Therefore, if Oscar believes in S2 what he believes in S1, Oscar believes in S2 that he has arthritis in his thigh. In that case, however, he believes *falsely* in S2 that he has arthritis in his thigh, since in S2 he does *not* have arthritis in his thigh. But there is no reason whatsoever to impute error to Oscar in S2. In S2, he is using the word 'arthritis' correctly; it does apply to the ailment in his thigh. Since Oscar does not believe falsely in S2 that he has arthritis in his thigh, Oscar does not believe in S2 that he has arthritis in his thigh.[5] Therefore, in S1 but not in S2 Oscar believes that he has arthritis in his thigh, which is exactly Burge's point. Of course, the argument as just laid out uses explicit though fairly elementary deductive logic, which is untypical of everyday life. But it also makes essential use of the thought experiment, to establish the premises of the reasoning, in part by rather

easy applications of the psychological capacities underlying our everyday ascriptions of propositional attitudes. Despite the residual opposition of some philosophers with internalist commitments in the philosophy of mind, there is no good reason for scepticism about the argument.

The arthritis example also brings out one role for philosophical expertise in some thought experiments: in this case, broadly logical expertise acquired through training in logic, a form of philosophical expertise which even experimental philosophers seem willing to grant. Such expertise is relevant not only to constructing the explicit argument, but also to avoiding various confusions to which the folk may be vulnerable. For instance, if one is careless about the use-mention distinction, one may be tempted to think that in S2 Oscar *does* have arthritis in his thigh, because the *word* 'arthritis' as used in S2 does correctly apply to the ailment in Oscar's thigh. Some ordinary subjects may indeed give false verdicts on Burge's thought experiment as a result of such undergraduate errors. They warrant no more scepticism than other undergraduate errors do. Alas, however, not even a Ph.D. in philosophy *guarantees* immunity to use-mention confusions.[6]

5. Acts of Judging and Evidence

Machery (2011) assumes that the main evidence for the truth of the key judgment in a thought experiment is the act of judging itself (even if it is poor evidence). For example, the main evidence that in the Gödel-Schmidt case 'Gödel' refers to Gödel is that (some) subjects

judge that in the Gödel-Schmidt case 'Gödel' refers to Gödel. Is this epistemological claim correct?

Machery justifies his assumption by analogy with ordinary judgments: 'If I judge of an object that it is a chair, my judgment that it is a chair is evidence that it is a chair because I am reliable at sorting chairs from nonchairs' (2011, 194). This remark blurs a crucial distinction between two issues. First, is the act of making the judgment evidence for its truth from the standpoint of a third party? Second, is the act of making the judgment evidence on which that very judgment is based? Clearly, these two questions can have different answers. Suppose that initially I know nothing about an object *o* except that there is such an object. I have the background information that Machery is reliable at sorting chairs from non-chairs. Now I learn just that Machery judges that *o* is a chair. Obviously, the probability that *o* is a chair on my evidence goes up considerably. In that sense, Machery's act of judging that *o* is a chair can of course be evidence for me that *o* is a chair. But that does not mean that his act of judging was evidence on which that very judgment of his was originally based. It could not have been, for his act of judging was not available as evidence until the judgment had already been made. Typically, he knows that *o* is a chair much more directly, by *seeing* that *o* is a chair. If he needs further evidence, he has much better and more direct evidence from perception: he can see that *o* has legs, a seat, a back, and so on. Even when *o* is no longer in sight, he can remember that *o* has legs, a seat, and a back. For Machery to go instead by the fact that he once judged that *o* was a chair would be a pointlessly indirect detour. And if for some reason he starts doubting that his original judgment that *o* was a chair was correct, the consideration that he did indeed make that judgment is unlikely to reassure him. It is unclear why anyone would attribute a special

evidential role to the fact of judging itself, except under the influence of the psychologization of evidence, which I have criticized elsewhere (Williamson 2007, 234-8).

Parallel considerations apply to thought experiments. Suppose that initially I know nothing about a situation GS except that there is such a counterfactual situation. I have the background information that Kripke is reliable at doing thought experiments. Now I learn just that Kripke judges that in GS 'Gödel' refers to Gödel. Obviously, the probability that in GS 'Gödel' refers to Gödel on my evidence goes up considerably. In that sense, Kripke's act of judging that in GS 'Gödel' refers to Gödel can of course be evidence for me that in GS 'Gödel' refers to Gödel. But that does not mean that his act of judging was evidence on which that very judgment of his was originally based. It could not have been, for his act of judging was not available as evidence until the judgment had already been made. Presumably, Kripke knows that in GS 'Gödel' refers to Gödel much more directly, by considering GS appropriately in his imagination. If he needs further evidence, he has much better and more direct evidence from noting the stipulated features of GS itself: he knows that in GS there is a stipulated historical connection of a certain kind between 'Gödel' and Gödel (which is good *evidence* that the former refers to the latter on any reasonable theory of reference for proper names). For Kripke later to go instead by the fact that he once judged that in GS 'Gödel' refers to Gödel would be a pointlessly indirect detour. And if for some reason he starts doubting that his original judgment that in GS 'Gödel' refers to Gödel was correct, the consideration that he did indeed make that judgment is unlikely to reassure him. Again, it is unclear why anyone would attribute a special evidential role to the fact of judging itself, except under the influence of the psychologization of evidence.

According to Machery (2011, 194 n4): 'it is hard to see what other kind of evidence [than the act of judging] could be put forward to support the claim that, e.g., in the situation described by the Gödel case "Gödel" refers to Gödel'. This incomprehension seems to be related to the error, against which section 2 warned, of regarding the crucial judgments in thought experiments as involving no role for ordinary evidence, which comes of forgetting how those judgments correspond to evidence-based judgments about observed cases.[7]

6. Error-fragility

The use of elaborate imaginary cases *is* a distinctive methodological feature of much contemporary philosophy, even though our verdicts on them do not form a psychological kind. Despite all that has been said, we might still reasonably hope for some independent corroboration of those verdicts. Even when verdicts on many different thought experiments corroborate each other, we might still hope for some independent corroboration of the lot of them. One can take that view while regarding the method of thought experiments as evidentially quite respectable. Compare Whewell's idea of the consilience of inductions: a conclusion supported by one sort of inductive evidence is much better off if it is supported by other sorts of inductive evidence too.

Still, if thought experimentation can yield *knowledge* of a fact, why should more support be needed? That is like asking: if naked-eye vision can yield knowledge of a fact, why should more support be needed? Methodological questions are not just about the

epistemology of a one-off situation. They concern what general epistemic policies we should follow, for instance in philosophy. Although naked-eye vision without further checks can yield knowledge, a general policy of relying on naked-eye vision without further checks must be expected to yield errors too, since the faculties we use in naked-eye vision are fallible. Similarly, although thought experimentation without further checks can yield knowledge, a general policy of relying on thought experimentation without further checks must be expected to yield errors too, since the faculties we use in thought experimentation are fallible.

The point is reinforced by what Alexander and Weinberg (2014) call 'error-fragility'. A method is error-fragile if it multiplies error: pursuing it tends to make one error produce many more. Pure deduction is an error-fragile method. Although genuine deductions preserve truth, an imperfect logician applying a purely deductive method will occasionally mistake fallacies for genuine deductions, with potentially disastrous consequences. By contrast, simple induction is not very error-fragile, when based on more or less independent observations. Requiring a consilience of inductions makes it even less error-fragile. Pure falsificationist methods are also error-fragile, since they involve rejecting a theory on the basis of a single counterexample. If the supposed counterexample is erroneous, one may reject a true theory. But analytic philosophers have typically used thought experiments in applying just such a falsificationist method. For instance, a proposed analysis of knowledge is rejected when one thought experiment is judged to yield a counterexample. Thus a single erroneous verdict on a thought experiment might eliminate the true analysis of knowledge (if there were one).

Evidently, we need some system of checks on thought experiments. That does not imply their marginalisation. After all, mathematics has an adequate system of checks on the error-fragile method of deduction without marginalising it at all. One mathematician's proof is checked by others, and in the long run even if a fallacy in the proof passes unnoticed a false 'theorem' is likely to be found incompatible with true ones. To some degree, a method based mainly on thought experiments has analogues of those error-correcting mechanisms. But does it is have them to a high enough degree? We might reasonably hope for a more robust philosophical methodology where the method of falsification by thought experiment is checked and balanced by other methods. But which other methods should they be?

Experimental philosophers will of course propose experimental methods. For these purposes, it does not matter whether philosophers were involved in designing and conducting the experiments. As noted in section 1, experimental science already has an important input to several branches of philosophy. But it is unclear how much it can offer to constructive theorizing in those branches where the experimental critique of thought experiments has been most salient, especially moral philosophy and epistemology. Results about what lay people think about goodness or knowledge is only very indirect evidence about which theory of goodness or knowledge is true. Nevertheless, it is not unlikely that received verdicts on some thought experiments *do* reflect cognitive bias of some kind, for instance when high stakes are involved, and we may hope that, in the long run, experimental methods will help us filter out such cases. And, of course, cognitive psychology will surely contribute much to epistemology through experimental studies of perception, memory, and reasoning, although one must not imagine that popularizing such work is an adequate substitute for properly epistemological theorizing.[8]

Some branches of philosophy, such as philosophical logic, have far more to gain from formal methods than from experimental ones.[9] We should not assume that moral philosophy and epistemology are nothing like that. Moral philosophy learns from mathematical decision theory and game theory. Epistemology learns from probability theory and epistemic logic. Of course, moral philosophy and epistemology cannot be reduced to branches of mathematics, on pain of losing their connection to their subject matter. Formal models of moral or epistemic phenomena need informal motivation. Nevertheless, they provide a powerful means for thinking through the consequences of moral and epistemological hypotheses.

Combining the use of mathematical models, results from cognitive psychology, and pre-theoretic verdicts on real or imaginary cases constitutes a more robust methodology than reliance on any one or two of those three sources. Each source can alert us to errors made through reliance on the others. A consilience of them gives us more robust grounds for confidence. For instance, mathematical modelling supports the conclusion of Gettier's thought experiments (Williamson 2013). Moreover, information from those sources must be integrated within the overall setting of informal philosophical theorizing in a broadly abductive spirit, where theories are compared by familiar criteria such as simplicity, strength, unifying power, and fit with the evidence.

What happens if we delete the pre-theoretic verdicts on cases from such a methodology? Suppose that we are interested in some philosophically central distinction that neither mathematics nor cognitive psychology themselves supply us with, such as the distinction between right and wrong or between knowledge and ignorance. Mathematics says nothing special about the distinction. Cognitive psychology may tell us how humans

apply it, but not whether they apply it correctly or incorrectly. If we want to start talking on our own behalf about the distinction, we must rely initially on our own pre-theoretic applications of it, even though we reserve the right to revise them in the light of subsequent theorizing. If we are not allowed to start from our pre-theoretic judgments about cases, then all we have left are our pre-theoretic *general* judgments about the distinction ('Ought implies can'; 'Knowledge implies belief'). But if we do not trust our particular judgments about the distinction, why trust our more general ones? After all, any pressure in the history of our species to apply the distinction correctly is far more likely to have come from the practical need to classify particular cases at hand correctly than from the theoretical desirability of formulating true generalizations about it. 'Stick to generalities' and 'Avoid examples' are not recipes for good philosophizing, or indeed good theorizing of any kind. Philosophy cannot be reduced to psychology; no clear or plausible picture of an alternative philosophical method has emerged from experimental philosophers' critique of armchair philosophy. There may indeed be a role for experimental philosophy in refining current philosophical method, but only once the method of experimental philosophy has itself been considerably refined.[10]

Notes

1 Arguably, what most epistemologists call ‘justified belief’ is better classified as *blameless* belief (Williamson 201X), but the experimental critique of Gettier cases concerns the denial of ‘know’, not the application ‘justified’, which most epistemologists use as a theoretical term, since they intend a restriction to *epistemic* (as opposed to moral or pragmatic) justification. In this chapter, I apply the term ‘justified’ in the way analytic epistemologists have usually done.

2 The complaint that the counterexamples in the text differ from philosophical thought experiments in being clear cases assumes what the negative program is trying to establish. By current philosophical standards, Gettier cases are clear cases of not knowing.

3 Note that the problem of the comparison class here primarily concerns the application of ‘usual’, not the application of ‘reliable’.

4 Machery makes these claims after considering cases involving the apportionment of blame, but does not restrict his claim to such cases.

5 A few loose ends need to be tied up, for example to ensure that Oscar in S2 does not have some other word that refers to arthritis and so does not apply to the ailment in Oscar's thigh. They do not affect the point in the text.

6 See Machery 2011, 206-12, and 2015, Williamson 2011, and references therein, for discussion of a more general defence of philosophical thought experiments by appeal to the phenomenon of philosophical expertise. I have not focussed on this defence here for two reasons. First, many of the issues it raises are specific experimental ones of the sort with which this chapter is not concerned. Second, most of the arguments from experimental philosophy discussed in this chapter can be rebutted without appeal to the phenomenon of philosophical expertise.

7 That one may have both direct evidence for a proposition by perception or imagination and also indirect evidence for it by knowing that others believe it does not undermine the points in the text.

8 We should also remember that the interpretation of real life experiments can involve cognitive bias of its own, such as concentration on those experiments that give the results one is hoping for.

9 Of course, experimental methods may show that many people are willing to assent to "It is and it isn't" when they feel pulled both ways about whether a borderline shade is red. That is roughly as much of a threat to classical logic as experimental evidence that many people are willing to assent to "One plus one equals one" when drops of water coalesce or "One plus one equals ten" when rabbits breed is to standard arithmetic. This is not to deny that there are connections between philosophical logic and the semantics of natural languages (for instance, in the study of conditionals), and that experimental methods are in principle relevant to the latter. Nevertheless, interpreted logical theories are not metalinguistic theories unless they happen to concern metalinguistic logical constants (such as a truth predicate), still less psychological theories. The appropriate methodology for testing them is similar to that for testing interpreted theories in mathematics, for instance set theories. See CHAPTER YY (EXPERIMENTAL PHILOSOPHICAL LOGIC) and CHAPTER ZZ (EXPERIMENTAL PHILOSOPHY MEETS FORMAL EPISTEMOLOGY) for more sympathetic accounts of the role of experimental methods in these areas.

10 Thanks to an audience in Oxford for discussion and to Joshua Alexander, Wesley Buckwalter, Joshua Knobe, Edouard Machery, Peter Millican, Jennifer Nagel, Justin Sytsma, and Jonathan Weinberg, for detailed written comments on earlier drafts of this chapter.

References

Alexander, Joshua. 2012. *Experimental Philosophy: An Introduction*. Cambridge: Polity Press.

Alexander, Joshua, and Jonathan Weinberg. 2007. "Analytic Epistemology and Experimental Philosophy." *Philosophy Compass*, 2: 56-80. DOI: 10.1111/j.1747-9991.2006.00048.x

Alexander, Joshua, and Jonathan M. Weinberg. 2014. "The 'Unreliability' of Epistemic Intuitions." In *Current Controversies in Experimental Philosophy*, edited by Edouard Machery and Elizabeth O'Neill, 000-000. London: Routledge.

Burge, Tyler. 1979. "Individualism and the Mental." *Midwest Studies in Philosophy*, 4: 73-121. DOI: 10.1111/j.1475-4975.1979.tb00374.x

Cappelen, Herman. 2012. *Philosophy without Intuitions*. Oxford: Oxford University Press.

Deutsch, Max. 2009. "Experimental Philosophy and the Theory of Reference." *Mind and Language*, 24: 445-466. DOI: 10.1111/j.1468-0017.2009.01370.x

Kripke, Saul. 1980. *Naming and Necessity*. Oxford: Blackwell.

Machery, Edouard. 2011. "Thought Experiments and Philosophical Knowledge." *Metaphilosophy*, 42: 191-214. DOI: 10.1111/j.1467-9973.2011.01700.x

Machery, Edouard. 2015. "Illusions of expertise". To appear in *Experimental Philosophy, Rationalism, and Naturalism: Rethinking Philosophical Method*, edited by Eugen Fischer and John Collins, 000-000. London: Routledge.

Mercier, Hugo, and Dan Sperber. 2009. “Intuitive and Reflective Inferences.” In *In Two Minds: Dual Processes and Beyond*, edited by Jonathan Evans and Keith Frankish, 149-170. Oxford: Oxford University Press.

Nagel, Jennifer. 2012. “Intuitions and Experiments: A Defense of the Case Method in Epistemology.” *Philosophy and Phenomenological Research*, 85: 495-527. DOI: 10.1111/j.1933-1592.2012.00634.x

Rose, David, and David Danks. 201X. “Turning Mountains Back into Molehills: In Defense of a Broad Conception of Experimental Philosophy.” *Metaphilosophy*, XX: XXX-XXX.

Sosa, Ernest. 2007. “Experimental Philosophy and Philosophical Intuitions.” *Philosophical Studies*, 132: 99-107. DOI: 10.1007/s11098-006-9050-3.

Weinberg, Jonathan, Shaun Nichols, and Stephen Stich. 2001. “Normativity and Epistemic Intuitions.” *Philosophical Topics*, 29: 429-460. DOI: 10.5840/philtopics2001291/217.

Williamson, Timothy. 2007. *The Philosophy of Philosophy*. Oxford: Wiley-Blackwell.

Williamson, Timothy. 2011. “Philosophical Expertise and the Burden of Proof.” *Metaphilosophy*, 42: 215-229. DOI: 10.1111/j.1467-9973.2011.01685.x

Williamson, Timothy. 2013. “Gettier Cases in Epistemic Logic.” *Inquiry*, 56: 1-14. DOI: 10.1080/0020174X.2013.775010

Williamson, Timothy. 201X. “Justifications, Excuses, and Sceptical Scenarios”. In *The New Evil Demon: New Essays on Knowledge, Justification, and Rationality*, edited by

Fabian Dorsch and Julien Dutant, forthcoming.