

# On the Limitations and Criticism of Experimental Philosophy

Theodore Bach

(forthcoming in Bauer, M. & Kornmesser, S. (eds.) Compact Compendium of Experimental  
Philosophy, De Gruyter)

Abstract: Experimental philosophy involves subjecting philosophical methods and judgments to empirical scrutiny. I begin by exploring conceptual, confirmational, and empirical factors that limit the significance of experiment-based and survey-based approaches to the evaluation of philosophical epistemic activities. I then consider specific criticisms of experimental philosophy: its experimental conditions lack ecological validity; it wrongly assumes that philosophers rely on psychologized data; it overlooks the reflective and social elements of philosophical case analysis; it misconstrues the importance of both procedural and evaluative forms of philosophical expertise; it incorrectly views psychological bias as incompatible with reliability; and it generalizes to a global, self-defeating skepticism about case judgment. I explain why these criticisms should be understood as converging and interdependent. I also set out a three-level model of philosophical case judgment that frames the criticisms.

## **1. Introduction: Limitations versus criticisms**

Here are three factors that limit the performance of my decade-old station wagon: the earth's gravity limits its vertical mobility, an outdated audio system limits listening choices to compact discs and radio, and a weak engine limits its ability to accelerate quickly onto highways. Some of these limitations are grounds for justified critique (the third, which is a design flaw), and some

are not (surely the first). The example suggests a rough but serviceable account of the difference between an item's or method's limitations, on the one hand, and grounds for its (justified) critique, on the other. A limitation indicates something that an item or method cannot do. Some limitations are more or less permanent (the first above), and some are not (the second). Some deserve scrutiny (the third), and some are uninteresting (the first). A criticism, on the other hand, is an evaluative claim about an item or method that is made relative to what that item is designed to do, what someone purports it can do, or more generally some type of error with respect to that item's or method's output. Not all criticisms are justified.

After discussing the boundaries and aims of experimental philosophy, I sketch general limitations (Section 3) and general criticisms (Section 4) so understood of the epistemic activity of experimental philosophy, targeting mostly the interesting and justified varieties.

## **2. Experimental Philosophy: what it is and aims to do**

We first need a handle on what experimental philosophy (commonly dubbed 'X-Phi') is and aims to do. The matter is not simple.<sup>1</sup>

X-Phi overviews typically distinguish a broad and narrow conception. Rose and Danks (2013) advance the former, claiming that X-Phi is "philosophical work that uses various empirical results, particularly from the cognitive sciences, in philosophical theorizing" (p. 515), with the further condition that the experiments "happen to have been done by the same individual[s]" (p. 515) who are doing the philosophical theorizing.<sup>2</sup> This counts the epistemic

---

<sup>1</sup> See Fischer, this volume.

<sup>2</sup> Stich and Tobia (2016) advance a similarly broad conception, but they drop Rose and Dank's "colocation" condition: "experimental philosophy is empirical work undertaken with the goal of contributing to a philosophical

projects of Piaget, Helmholtz, and Kohlberg as instances of X-Phi. Knobe and Nichols (2017) also invoke a broad conception, claiming that X-Phi “brings together” the questions and frameworks of traditional philosophy, on the one hand, and “the kinds of experimental methods traditionally associated with psychology and [the] cognitive [and social] science[s]” (p. 1), on the other. This counts Fodor, Crick and Koch, Ruth Millikan, and nearly all philosophical naturalists as experimental philosophers.

The narrow conception understands X-Phi as the application of experimental results (generally survey-centered studies) to the evidential use of philosophical intuition or, more generally, the dialectical use of “the method of cases” (see below, Section 4.2). Most but not all instances of X-Phi fall into this category.<sup>3</sup>

With respect to the aims of X-Phi, it is standard to distinguish “positive” and “negative” programs.<sup>4</sup> The positive program aims to make progress on the questions that philosophers traditionally seek to answer, e.g., *what is knowledge?* Advocates of this program claim that if eliciting intuitive verdicts about cases is useful for adjudicating competing philosophical models, then it is better to canvas intuitions empirically than to rely on armchair assumptions about (say) their uniformity (see, e.g., Knobe 2003, Nahmias et al. 2006). In contrast, the negative program emphasizes that intuitions, or more generally case judgments, are often steered by epistemically

---

debate [...] sometimes people doing experimental philosophy conduct the experiments, but sometimes they don't” (p. 5).

<sup>3</sup> Paradigmatic examples include survey-centered studies that suggest philosophical case judgments (about knowledge possession, to pick one of many topics) vary with the order of case presentation (Swain et al. 2008), the font of case presentation (Weinberg et al. 2012), and the cultural background of the person making the judgment (Machery et al. 2004). Readers should consult Machery (2017, Chapter 2) and the chapters in this volume for an overview of the experimental findings.

<sup>4</sup> Some commentators mention a third program, which is (very roughly) to advance our understanding of psychological explananda, for example mechanisms of belief-formation.

irrelevant factors. In light of this influence, advocates of the negative program claim that philosophers should not rely uncritically on intuitions as sources of evidence. The kernel of the negative program is “the Restrictionist Challenge” (Alexander and Weinberg 2007, Weinberg et al. 2010): until intuition-users can show that they have corrected their evidential use of intuition to sufficiently insulate them from non-truth tracking influences – or until they can show that the intuitions were never so influenced in the first place – they ought to stop using the intuitions evidentially in the error-prone contexts.

There is a clear difference in emphasis between the negative and positive program, but the distinction may also be misleading. Those aligned with the negative program are just as interested in, and believe X-Phi is instrumental to, answering philosophical questions accurately as those aligned with the positive program. Moreover, the negative program is not committed to the claim that the method of cases or the evidential use of philosophical intuitions is necessarily or universally epistemically unreliable (Weinberg 2017).<sup>5</sup>

At this point it is fair to ask whether unclarity about X-Phi’s boundaries and aims could itself mark an important limitation on X-Phi’s capacity to reform first-order philosophy or generate metaphilosophical insight. If there is a limitation or grounds for criticism here, then given similar doctrinal debates in nearly all philosophical sub-disciplines (especially so given the recent surge in metaphilosophical discussion), it would generalize broadly. It would be a limitation of philosophical discourse generally, not a limitation of X-Phi specifically.<sup>6</sup>

---

<sup>5</sup> But see the “radical restrictionism” advocated for in Machery (2017), though Machery’s restrictionism applies only to cases that elicit modal philosophical claims.

<sup>6</sup> One is reminded of complaints in the 1980’s and 1990’s that philosophical feminism lacked doctrinal unity or that its subject matter (the category *woman*) was ill-defined (see, e.g., Young 1994). These types of concerns are generally worth exploring, but they are also reflections of ongoing disagreements about deeply theoretical issues concerning classification, reference, conceptual change, natural kinds, and so forth. I do not think that experimental

But perhaps a stronger case can be made that there is a natural kind of epistemic project to which many (but not all) self-claimed instances of X-Phi belong. I submit that a methodological or metaphilosophical conception of X-Phi deserves consideration here. On this conception, X-Phi's essential feature is its application of experimental results to philosophical methodology – to whatever psychological, social, and even biological mechanisms underwrite philosophical output, whether that output is a private judgment about a philosophical case or a trend in professional philosophers' credences. This conception encompasses paradigmatic, survey-based studies of philosophers' method of cases. Less obviously, it counts Kathy Davis's (2008) "Intersectionality as Buzzword" as X-Phi, as that article applies Murray Davis's (1971, 1986) empirical investigations of the properties that causally explain an academic paper's professional uptake to philosophy. Contrary to the broad conception, it does not count Helmholtz, Crick and Koch, Piaget or Millikan as conducting X-Phi, as these researchers are not bringing to bear experimental results specifically on philosophical methodology (even if these researchers are addressing philosophical questions in the spirit or empirically informed philosophy).

I will assume this methodological conception of X-Phi in what follows, with a focus on paradigmatic X-Phi (i.e., the narrow conception).

### **3. Three Limitations of Experimental Philosophy**

Here I discuss three factors that limit the type and strength of evaluative claim about philosophical method that is warranted given the empirical findings of X-Phi. These discussions

---

philosophers (or philosophical feminists) should feel any more hamstrung about these doctrinal debates than evolutionary biologists should feel hamstrung about the ever-increasing number of species concepts (currently around 27). See also Boyd (2021, p. 2874).

are not advanced here as criticisms. However, and as will become clear below, it is possible to leverage these limitations in support of specific lines of X-Phi critique.

### 3.1 The central limitation of experimental philosophy

We generally do not have access to – or empirical knowledge of – which specific material conditions can in principle, let alone in practice, distinguish the correctness of two rival but empirically adequate philosophical models. For example, when philosophers disagree about whether all human motivation reduces to self-directed desires, or whether free will requires an indeterministic world, or whether the possession of meaningful mental states requires a certain type of history, there is no empirical state of affairs the observation of which will tell us who is correct and who is incorrect.<sup>7</sup> And while it is tempting here to appeal to episodes of agreement among professional philosophers as a capable substitute for the empirical confirmation of philosophical accuracy (at least for that small subset of philosophical models for which there is something like a consensus), there is (perhaps ironically) disagreement about the merits of such a method (see, e.g., Kelly 2016). Moreover, the history of philosophical consensus has a worrisome track-record.<sup>8</sup>

---

<sup>7</sup> For discussion, see Williamson (2007), Sosa (2011), Papineau (2011), Paul (2012), Ryberg (2013), Nolan (2015), Ludwig (2018) and Bach (2019, 2021). Note that this confirmational limit on the assessment of philosophical models applies regardless of whether the disagreeing parties espouse material (synthetic) or semantic (analytic) aims when employing philosophical case analysis. (See Section 4.1 below for further discussion of the distinction between these aims). Note that it also applies to many disagreements about which facts obtain in philosophical thought experiments.

<sup>8</sup> As Lycan (2017) remarks: “There are periods of very wide agreement, but they are pathetically short and geographically local. Just in my lifetime, we Anglo-Americans have seen sense-datum theory reign and then be scorned, mind-body materialism reign and then come under heavy attack, and so forth. The late Jerry Katz once predicted to me that Moorean moral intuitionism would make a big comeback; and I would never doubt that

This confirmational predicament limits how effective the empirical methods of experimental philosophy can be for indicating whether particular influences on the construction and judgment of philosophical models are truth-tracking or non-truth-tracking. For example, with respect to the X-Phi negative program, while experimental philosophers might demonstrate that a particular factor (order of case presentation, cultural background, etc.) exerts a causal influence on philosophical judgment, they are far more limited in their ability to demonstrate that this influence is epistemically “distorting.” To make this stronger claim about epistemic distortion, one needs to add a metaphilosophical assumption (e.g., that the influence of cultural background on philosophical case judgment is non-truth-tracking) the truth of which is difficult to confirm empirically. And if one is able to secure such a stronger claim, one still needs to show that the effect of the distorting factor is significant – that its size and relationship to philosophers’ case analysis renders that judgment-type, or its underlying mechanism, unreliable (see Section 4.4 below).<sup>9</sup>

To put this point into focus, it is helpful to consider how the *absence* of similar confirmational limitations in *non*-philosophical domains permits evaluators of those domains more effective use of empirical methods for performance assessment. In fact, there is an extensive empirical literature on the accuracy of professional judgment (including forms of judgment that many philosophers will classify as intuition-based) for a wide variety of non-philosophical domains (chess, clinical psychology, finance, livestock judging, and so on). A central and surprising finding is that in many domains, credentialed experts often fail to

---

idealism will too, possibly still within my lifetime. Philosophical consensus is far more the result of Zeitgeist, fad, fashion, and careerism than of accumulation of probative argument” (p. 109f.).

<sup>9</sup> Nor can one resolve this issue simply by generalizing criteria for effect significance from non-philosophical experimental contexts (but see Weinberg 2016 and Machery 2017 for attempts to address this concern).

outperform both novices and statistical models.<sup>10</sup> Putting aside the question of *why* this is (but see below), we should ask: *how* are empirical researchers able to make these determinations? The answer, in brief, is that the researchers have a clear and quantifiable success metric relative to which they can assess the accuracy of (non-philosophical) expert judgment and methodology. Financial advisors predict the market will turn bullish, clinicians judge that intervening with medication *y* will bring about improvement *z*, parole experts predict that inmates with property *x* are the most likely to reoffend, and so forth. In all these cases, there is a (future) fact of the matter that indicates whether the expert's judgment – as a prediction, intervention, or rule-following consideration – was correct or incorrect. The auditors' epistemic access to this metric – their knowledge of the truth-determining material events – is what allows them to assess accurately the epistemic quality of the experts' judgments.

But as discussed above, in the philosophical case, we do not have access to such a metric. For example, we do not know which specific material conditions can in principle distinguish the correctness of competing but empirically equivalent philosophical models of causation. As a result, we are limited in our ability to know which causal influences on philosophical model-building and judgment are epistemically distorting versus epistemically relevant.

We should not take this point too far. Some assumptions about epistemically distorting influences are warranted regardless of whether we can empirically confirm the accuracy or inaccuracy of theoretical models the construction of which were subject to those influences. For example, surely the influence of font-type (Weinberg et al. 2012) or racially coded name changes (Uhlmann et al. 2009) on case judgment does not reliably steer towards greater accuracy, and

---

<sup>10</sup> See, e.g., Shanteau (1992). These domains include finance, criminal justice, clinical psychology, graduate admissions, and political forecasting.



this is something that we can affirm prior to knowing which philosophical models are most accurate. The concern, however, is that the range of causal influences to which that type of warrant applies may be quite limited. For example, many researchers had assumed that the influence of presentation order on philosophical case judgment must be epistemically distorting (non-truth-tracking), but as Horne and Livengood (2017) point out, there are grounds for claiming that ordering effects can reveal an epistemically virtuous form of “updating” (that is, learning). Similarly, Ludwig (2010, 2018) has pointed out that the influence of cultural background on philosophical case judgment about the reference of proper names is not epistemically irrelevant if different cultural backgrounds underwrite different exposures to factors that inform accurate versus inaccurate views about the nature of reference.<sup>11</sup> Even with respect to font-type, a case can be made that its influence on judgment is epistemically relevant (particularly if the font type is hard to read and thus more likely to induce deeper thinking).<sup>12</sup>

It is not uncommon for experimental philosophers to respond to such proposals by requesting empirical support for the purported epistemic relevance of the influencing factor (see, e.g., Weinberg et al. 2010, Alexander 2012, Sytsma and Livengood 2015, Machery 2017). If the relevant types of empirical tests are available, then this is a sensible request. The problem – and this stems from the central limitation discussed here – is that often such tests are not available.

---

<sup>11</sup> In Ludwig’s (2018) terms: “all that can be shown by these results is that not everyone is getting it right, and that there can be factors that correlate with different cultural, ethnic, or socio-economic backgrounds that can contribute to errors” (p. 391). Some restrictionists charge this type of suggestion with “chauvinism” or “ethnocentrism” (e.g., Machery 2017, p. 106). It is not clear that the charge is warranted. The claim about epistemic relevance does not require that “we” are getting it more correct than “them” – it only requires that some group is getting it more correct. If there *are* reasons to privilege the view of a particular group – perhaps one of the groups has received expert training – then one might side with Jackson’s (2011) suggestion that “talk of chauvinism is a misdirection” in this context (p. 469). I discuss some of these issues below in the context of the expertise defense (see Section 4.3.3).

<sup>12</sup> See Weinberg et al. (2012, p. 218, fn. 22).

For example, if researchers do not know in advance what philosophical model of reference is correct, then it is unclear how to isolate via empirical tools subtle developmental differences that might causally explain why one group's judgments are slightly more aligned with the correct model (whatever that model is). We should certainly be on guard against post-hoc or "just-so" stories about the epistemic virtue of a causal influence on philosophical judgment. But we should also be careful not to weaponize dialectically, sometimes through verificationist claims about burdens of proof, the confirmational predicament of much philosophical discourse.

Some experimental philosophers invoke empirical data on non-philosophical expert performance to bridge (or rather, circumvent) the impasse generated by the confirmational limitation discussed above. The basic idea is to identify the features that causally explain poor versus superior expert performance in non-philosophical domains (domains for which the confirmational limitation does not apply), and then investigate whether those features manifest in philosophy (see especially Weinberg et al. 2010, Clarke 2013, Ryberg 2013). We know that, in the case of many non-philosophical domains, exposure during training and job experience to direct and environmental learning feedback about the quality of one's judgments leads to superior performance (and its absence leads to inferior performance). Is that type of direct and environmental learning feedback available to philosophers? Experimental philosophers provide reasons – reasons connected to the confirmation issue discussed throughout this section – for why it is not.<sup>13</sup> They then reason inductively: given this shared developmental condition between philosophy and the poor-performing non-philosophical expert domains, we should doubt (even if

---

<sup>13</sup> Clarke (2013) is especially clear on this point: "In the case of philosophical intuitions, however, direct feedback from the environment is typically unavailable. We cannot directly discover what knowledge really is or what morality really demands" (p. 193).

we cannot empirically confirm it) whether expert philosophical training and experience promotes more accurate theorizing and case judgment.

This is a good example of experimental philosophy – the use of empirical methods and data to evaluate philosophical methodology. Notably, it does not rely on surveys. Nonetheless, we should raise some objections. First, the types of non-philosophical judgment that according to the empirical literature require direct and environmental learning feedback for superior performance are generally rule-following, predictive, or intervention-based judgments. However, these aims and types of judgment content are unrepresentative of philosophical theories and judgment. Philosophical theories and judgments typically have unificatory and explanatory aims rather than predictive or rule-following aims. If that is correct, we should reject generalizations to philosophy that are drawn from data about professional domains that have non-representative goals and success conditions. Second and related, if our goal is to locate an empirically studied non-philosophical base from which to generalize to philosophy, and if we take seriously the abstract unificatory aims of much philosophical theorizing (e.g., Paul 2012), then we are led to areas of the empirical literature on expertise that are much more encouraging for expert philosophers. Empirical studies indicate that experts in feedback-deficient environments demonstrate superior performance with respect to relational (as opposed to superficial) retrieval, simulation, ability to discover important causal-explanatory relationships and categories, and ability to grasp the meaning and importance of unusual events (see Bach 2021).

### 3.2 Empirical Limitations of Experimental Philosophy

The previous section examined how the distinctive targets and aims of philosophical analysis impose limits on what we can learn about the epistemic success of philosophical analysis based

on empirical investigation of its underlying methods. This section sketches limitations sourced in the specific empirical methods commonly employed by experimental philosophers.

An immediate concern is that X-Phi's adoption of the experimental methods of the social sciences will inherit whatever limitations apply to those methods generally. One limiting factor – replication failure – is especially salient here (see, e.g., Woolfolk 2013, Seyedsayamdost 2015). Some influential X-Phi findings, for example gender effects on semantic judgments (Buckwalter and Stich 2014) and cultural background effects on epistemic judgments (Weinberg, Nichols, and Stich 2001), have failed to replicate. Many other influential findings have been reproduced. Researchers participating in the “XPhi Replicability Project” (Cova et al. 2021) provide evidence that suggests there is greater replication of X-Phi results than there is of results in social psychology generally (see also Stuart et al. 2019, Colombo et al. 2018, Machery 2017).

If we grant that many of the empirical findings of X-Phi are genuine and replicable, we should still inquire if there are features of survey-centered reports that limit their philosophical and metaphilosophical significance. One feature of concern is that they are sourced in self-report data. The goal of much X-Phi is to uncover epistemically irrelevant influences on “intuitions” or a type of cognitive judgment that is evidentially (or at least dialectically) instrumental to “the method of cases.” But given possible differences between what people say about philosophical cases and what they believe about philosophical cases, data restricted to the former provide a limited and potentially misleading understanding of the latter (Kauppinen 2007, Cullen 2010, Woolfolk 2013).

There are various ways that experimental philosophers have addressed this limitation, for example developing more carefully designed studies (e.g., incorporating pilot testing of subject comprehension). As Rose and Danks (2013) suggest, experimental philosophers can also support

survey self-report data with “behavioral measures, neuroimaging data, and other measures or cognitive functioning” (p. 525), although these additional tools come with their own, mostly logistical, limitations.

Regardless of the increasing sophistication of X-Phi studies, this limitation continues to inform several lines of criticism of both positive and negative X-Phi programs. Some critics allege that ambiguous vignettes and probe questions stoke verbal disagreement rather than reveal distorting influences (particularly with respect to vignettes about reference, which several critics have claimed are ambiguous between speaker’s reference and semantic reference – see Ludwig 2007, Deutsch 2009, and Heck 2018).<sup>14</sup> Sosa (2009, 2015) suggests that verbal disagreement, in addition to being caused by vignette ambiguity and contextual variation, results when survey participants draw from different conceptual backgrounds to fill in differently the missing details of a vignette (much like how people draw from their unique conceptual backgrounds to flesh out differently scenes from a novel). Other critics claim that X-Phi survey studies lack ecological validity. An experiment lacks ecological validity if it includes artificial conditions (e.g., asking human subjects to memorize a series of nonsense syllabus) that are importantly different from the part of the actual world that we are trying to understand (e.g., school learning) (Neisser 1976, p. 33).<sup>15</sup> According to these critics, X-Phi surveys misrepresent philosophers’ actual dialectical employment of the method of cases (see Section 4.2 below). Related to this, many critics claim that X-Phi surveys are inadequate for the purpose of informing us about the types of

---

<sup>14</sup> For example, in the case of X-Phi surveys that ask participants to say who John is talking about when John uses the name ‘Gödel’, participants might interpret this as asking about the language conventions for the referent of ‘Gödel’ (semantic reference) or about who John intends to use that name to talk about (speaker’s reference).

<sup>15</sup> As Neisser also points out, criticisms based on this notion lack force if they do not identify the specific real-world features that the experiment leaves out. In Sections 4.2 and 4.3, I specify the omitted real-world features that inform ecological validity-based criticisms of X-Phi.

philosophical case judgment that dialectically matter – “reflective” judgments and “expert” judgments (see Section 4.3.2 and Section 4.3.3 below).

### 3.3 Conceptual Limitations of Experimental Philosophy

A third and conditional limitation of specifically restrictionist X-Phi projects is the following: to the extent that restrictionist claims about philosophical methodology are warranted, one is thereby limited in which conceptual, philosophical resources are available for the construction of philosophical theories and arguments. If per the restrictionist challenge we ought to treat judgments elicited by various thought-experiments as epistemically suspect, then we ought to treat whichever philosophical concepts or theoretical frameworks centrally depend on those judgments as also epistemically suspect. This limits the pool of claims and theories from which we can construct philosophical arguments, including those that might inform the evaluative and interpretive claims of X-Phi.

The extent of this limitation will depend on at least two factors. First, it depends on the strength of the restrictionist claim. Machery’s radical restrictionism (see Machery 2017), which employs inductive arguments and metaphilosophical assumptions to restrict the use of all (modally ambitious) instances of the method of cases, cordons off more conceptual resources than moderate forms of restrictionism that target only judgment-types shown empirically to exhibit susceptibility to irrelevant influencers. Second, the extent of this limitation depends on whether there are alternative ways of supporting the conceptual resource. Here it is helpful to have a positive account of what makes evidential practices reliable and that does not recruit the problematic evidential appeals (see, e.g., Weinberg 2007; for discussion, see Section 4.5 below).

At any rate, it is not difficult to see how this limitation might be directed critically at X-Phi itself, making it a victim of its own success. Indeed, some critics claim that X-Phi is in this sense self-defeating, while others allege that it leads to a generalized and unacceptable form of judgment skepticism. I discuss these objections in Section 4.5.

#### **4. Specific Criticisms of Experimental Philosophy**

It has become standard to chunk up criticisms of X-Phi into distinct categories or “defenses” (e.g., Sytsma and Livengood 2015, Machery 2017): mischaracterization, expertise defense, reflection defense, and so on. I will be doing some of that here, but I think that the practice is misleading. It gives the impression that the criticisms are isolated – that the defender of X-Phi can deal with one type of criticism, then move on to deal with the next type, and so on. It obscures synergistic relationships between the criticisms. We would do well, I suggest, to think of X-Phi criticisms in terms of Duncker’s radiation problem (see Gick and Holyoak 1980): they are individual rays of radiation launched from different angles towards a central target, and while no single ray may be powerful enough on its own to eradicate the target, when considered as a converging network their overall destructive effect is considerably more compelling. I will flag examples of this converging effect as I move through the individual criticisms.

4.1 X-Phi findings are not relevant to philosophical methodology because philosophers do not rely on intuitions

Much early X-Phi resourced experimental findings to challenge philosophers’ evidential use of intuitions about cases. Several critics, particularly Williamson (2007), Deutsch (2009), and Capellen (2012), rejected that philosophers employ cases for the purpose of eliciting “intuitions”

– intellectually compelling seemings of some sort – that are supposed to serve an important evidential or argumentative role.<sup>16</sup> If these critics are correct, then empirical reports about distorting influences on philosophical intuitions would not seem relevant to actual philosophical argumentation and theorizing.

Both Deutsch (2009, 2015) and Capellen (2012) develop this criticism through detailed analysis of actual philosophical texts. Such textual scrutiny, they claim, reveals that philosophers’ psychological states (i.e., intuitions) do not play the evidential role that the restrictionists and positive experimental philosophers suppose. In Section 3.2, I mentioned that X-Phi surveys may be limited in their capacity to recreate the conditions of real-life philosophy. We can view Deutsch’s and Capellen’s text-based criticism in this light – as illustrating how a chasm between X-Phi lab conditions and real-life philosophy conditions permits a misleading picture of philosophical case analysis and the purported evidential role of intuitions.

According to Deutsch and Capellen, philosophers mostly construct arguments rather than rely on intuitions, and they use all manner of (non-psychological) data and methods to do so (see, e.g., Capellen 2012, p. 196). One standard method of argumentation, employed in both philosophical and non-philosophical contexts, is offering or ostending to a counterexample to challenge a generalization. Deutsch (2009) reports that Kripke’s discussion of the Gödel case does just this: it is “no different in principle from the method someone might use in arguing against the generalization that, for example, *all mushrooms are edible*, namely by pointing to a poisonous variety of mushroom” (p. 447). The point generalizes. Similarly, Williamson claims that in Gettier cases, it is simply a fact that there is not knowledge. For both Williamson and

---

<sup>16</sup> A related criticism was that experimentalists did not provide sufficient clarity about what their target is (i.e., intuitions). See especially Williamson (2011, 2013) on this point. See Capellen (2012, Chapter 7), Nado (2016), Pust (2019), and Horvath (this volume) for X-Phi-themed overviews of what intuitions are or are supposed to be.



Deutsch, then, it is an objective state-of-affairs (a poisonous mushroom, an opportunistic Gödel) and not some psychological feature of the apprehension of that state of affairs that grounds the counterexample.<sup>17</sup>

One might counter that it is the hypothetical nature of philosophical cases that makes their contemplation and classification as counterexamples intuition-based rather than fact-based. After all, these are thought-experiments that we conduct in our imaginations, so it would seem that any dialectically relevant “facts” will be psychological rather than material. In response, Williamson and others have pointed out that there are real-world analogues of many such cases (e.g., Gettier cases in which stopped clocks show the correct time). And if judgments about these actual cases are similarly construed as intuition-based, then negative X-Phi’s restrictionism appears to generalize into a global skepticism about the everyday (and scientific) judgment of actual cases (see Section 4.5).

But what about philosophers who explicitly describe themselves as appealing to special psychological states – intuitions – as a form of philosophical evidence, and who also provide explicit arguments for the epistemic importance of doing so (e.g., Bealer 1998, Goldman and Pust 1998, BonJour 1998)? These would be philosophers who, pace Capellen (2012), insist that their intuition talk is neither stylistic nor disposable scaffolding for non-intuition depending philosophical argument. And what about philosophers who claim that a central goal of philosophical theorizing and argumentation is probing the analytic commitments of one’s own concepts (e.g., Jackson 1998)? Aren’t at least *these* philosophers and their self-described methods worthy targets of both negative and positive X-Phi programs?

---

<sup>17</sup> Thus Williamson (2007, Chapter 7) cautions against the urge to psychologize philosophical evidence (an urge made stronger by the confirmational limitation discussed in Section 3.1). See Alexander (2010) for a contrary view.

The first thing to observe here is that, given the on-going trend of philosophical naturalism and empirically informed philosophy, this group of philosophers appears small and growing smaller.

Second, one can challenge whether these philosophers are correctly describing their own methods. This challenge can be made empirically by: (a) citing evidence that people are prone to confabulation and, more generally, are not very good at identifying their own cognitive processes, and (b) flagging that there is no current or foreseeable empirical evidence that warrants positing a special cognitive faculty dedicated to intuition. The challenge can also be made on conceptual grounds. Papineau (2009, 2011) offers compelling arguments that philosophers who claim to employ the case method for the purpose of making analytic judgments (e.g., about the application conditions of their own concepts) are, ultimately, invoking cases to make synthetic judgments.

Third, even if we were to grant that there is a faction of philosophers for whom it is descriptively accurate to characterize as relying evidentially on a richly-construed notion of philosophical intuition (or philosophers who truly employ the case method only for the purposes of conceptual analysis), then we might urge these philosophers to stop doing so (and for reasons not deriving from, but not incompatible with, negative program X-Phi findings).<sup>18</sup> See Millikan (2000), Papineau (2009), Kornblith (2014), and Boyd (2021) for compelling arguments in this direction.<sup>19</sup>

---

<sup>18</sup> In contrast, this position is incompatible with the aims of the positive X-Phi program.

<sup>19</sup> The three considerations set out above, along with the arguments made by Deutsch and Capellen, also apply to, and limit the significance of, X-Phi surveys (see Kuntz and Kuntz 2011; see Sytsma and Livengood 2015 for discussion) in which 23.5 percent of surveyed professional philosophers said that intuitions were essential for philosophical justification and half said that they were useful for philosophical justification.

4.2 X-Phi findings (and evaluative claims) are not relevant to philosophical methodology because they are based on surveys (and a conception of philosophical method) that overlook or mischaracterize important dialectical features of the “method of cases”

More recent X-Phi evaluative claims about philosophical practice do not assume that philosophers rely on psychologically unique events or states called “intuitions” as evidence. As Machery (2017) explains:

It is thus unfortunate that experimental philosophers, including myself, have followed the philosophical tradition in describing the method of cases as eliciting intuitions, and have given the impression that their argument was directed at the alleged use of intuitions in philosophy. It is not; the target is the method of cases. (p. 178)

On this construal, when philosophers evaluate cases, they deploy the same judgment-forming cognitive mechanisms that they and others use generally (Machery 2017, p. 21). The epistemic concern derives from a set of “disturbing” characteristics of the types of cases that philosophers tend to consider, for example the unusual and hypothetical nature of these cases. These disturbing characteristics purportedly corrupt philosopher’s judgments about what facts obtain in the target cases, and the evidence for that corruption is X-Phi’s findings of presentation and demographic effects on philosopher’s judgments about these cases (see Section 4.5).

Critics of X-Phi restrictionism allow that this construal’s omission of intuition-talk makes progress – its characterization of philosophical methodology is more closely aligned with what philosophers actually do, which is provide arguments and employ judgments about facts rather

than defer to psychologized forms of evidence called “intuitions.” Nonetheless, they insist that this “minimalist” conception of philosophical case judgment, and especially its attempted simulation in the lab by experimental philosophers (i.e., survey-studies), omits essential dialectical features of philosophers’ actual use of cases. Given this disparity, we are sharply limited in what evaluative inferences about actual philosophical practice we are warranted drawing from the experimental studies.

What are the omitted dialectical features? Critics point out that philosophers’ actual dialectical use of cases is socially embedded (Williamson 2007, 2013; Ludwig 2007, 2018; Kauppinen 2007, 2018). As Ludwig (2018) writes:

Philosophers aim to arrive at a reflective judgment about a case and then to review it in the light of other judgments (their own and others) and more general theoretical considerations. They do not simply record their spontaneous judgments and take the third person stance toward them as neutral observations to be explained. [...] we do not do this like hermits in the woods: we try out ideas and thought experiments on others, give and publish papers, take criticism, make revisions, try out new ideas generated in this process, and so on. (p. 388)

Critics point out that the cases considered by philosophers are eclectic, possessing non-generalizing domain-specific epistemic features (Capellen and Deutsch 2018, Capellen 2020). As Capellen (2020) puts it:

The term ‘method of cases’ doesn’t denote a theoretically useful class. It encourages the thought that there’s uniformity where there isn’t. The thing we call ‘cases’ come in too many varieties. The source of their degree of difficulty is multifaceted and cannot be separated from their specific subject matter. (p. 6)

They point out that the case method it is extended through time, both interpersonally and intrapersonally (e.g., Jackson 2011, Williamson 2009). Williamson (2009) says that the application of X-Phi data

ignores the difference between one-off individual judgments and consensus reached through the interaction of many participants in a public philosophical debate, conducted over several years in conferences and journals. (p. 474)

They point out that philosophers appeal to cases to reveal truths rather than cognitive judgments. Deutsch (2020) claims that the method of cases

does not involve everyday judgments as opposed to intuitive judgments. Rather, it does not involve judgments, period – not as a component of the method itself. (p. 769–770)

The overarching objection here is that, by overlooking these central features of the case method in both description and experimental simulation, experimental philosophers, including those that have dropped reference to intuition, continue to target “a mythical conception of ‘the method of

cases” (Kauppinen 2018, p. 69), a “caricature of philosophical method” (Ludwig 2018, p. 388), and “a fictional class” (Capellen 2020, p. 6).

#### 4.3 Philosophical Reflection and Expertise

The above criticism is a bridge to other, more fine-grained objections, particularly the criticisms that X-Phi fails to take sufficiently into account philosophers’ “reflective” and “expert” judgments. To unpack those specific criticisms, we need a richer, three-level model of the psychology and sociology of the method of cases to which we can appeal. Such a model needs to be empirically adequate, and it needs to make explicit the resources used by X-Phi critics who assert the importance of philosophical expertise and reflection.<sup>20</sup>

##### 4.3.1 *A Three Level Model of Philosophical Judgment*

At the bottom level (Level 1) are snap judgments about category membership. These might be unbidden, immediate, not fully conscious judgments concerning the classification of features of the target case. Still, these judgments do not emerge from an epistemic vacuum. They are sprung from some schema, theory, or other cognitively stored knowledge representations (see, e.g., Kahneman and Klein 2009, Kornblith 2007), and in that respect they are theory-laden. At the next level, Level 2, are considered, reflective judgments. Here, one thinks carefully about reasons for making the classificatory judgment. One might uncover to some extent the reasons that drove the snap judgment. One might sift through alternative reasons that lead to contrary judgments. One considers arguments, implications, and explanatory values of competing classificatory judgments. And so on. At the last level (Level 3), the reflective procedures of

---

<sup>20</sup> For a somewhat similar model of philosophical case judgment, see Strevens (2019, Section 4.3).

Level 2 are extended through time as well as the social, expert community. One talks with other philosophers, receives feedback from commentators and peer reviewers, reads articles and arguments, and so on. Given that Level 1 judgments are theory-laden, the fruits of Level 2 and Level 3 judgments feed back causally into Level 1 judgments, affirming or changing them in various ways. This is like how if one receives decades of formal chess training, then one's snap judgments about position classification will change as a result.

These judgment levels might function in a philosophical context as follows. When first confronting the Swampman thought experiment, one tokens the snap (Level 1) judgment that Swampman has contentful (meaningful) thoughts. But then (Level 2) one reflects more carefully on the case. One considers whether one's association between skilled behavior and underlying contentful cognition was the source of the snap judgment and, if so, whether that provides sufficient warrant for the classification. One considers analogous cases, for example that of a collection of twigs on a beach randomly wind-blown to form H-E-L-P but that does not carry the semantic information *help*. One is now inclined to override and revise the Level 1 snap judgment. At Level 3, one talks with other philosophers about Swampman. One writes about Swampman, receiving feedback from commentators and peer-reviewers. One reads the arguments from all the articles in the 1996 *Mind and Language* journal forum on Swampman. One considers again and affirms one's decision to revise the snap judgment.

#### 4.3.2 *Reflection*

The various references to “reflecting” and “considering” above, as well as some of the quotes from Section 4.2, make clear how an objection based on reflection will run (as well as why that objection is a species of the mischaracterization objection). On this view, any attempt to evaluate

or simulate the method of cases will fail to the extent that it misleadingly construes philosophers as epistemically beholden to Level 1 snap classificatory judgments. The very purpose of developing and socially disseminating philosophical cases, the objection continues, is to spur an extended process of careful and socially mediated reflection on competing classificatory judgments – a process that includes socially mediated reflection on different sets of arguments for the competing judgments, comparison of the explanatory and unificatory profiles of the competing classifications, and so on. Versions of this objection can be found in Ludwig (2007), Kauppinen (2007, 2018), Kornblith (2007), Williamson (2009), Cullen (2010), Bengson (2013), and Hannon (2018).

To address this appeal to reflective judgments, experimental philosophers have developed survey studies that they claim incorporate the reflective features. However, it is unclear whether X-Phi surveys can in principle, let alone in practice, probe the types of extended and socially mediated reflective judgment processes that according to critics are central to philosopher's actual use of cases (Kauppinen 2007). For example, it is unclear how experimental surveys can capture the three levels of judgment – and diachronic interactions between them – as sketched above. Kneer and colleagues (forthcoming), conceding that the reflection-based critique of X-Phi “has not yet received enough attention” (p. 1), attempt to simulate in surveys only what they term a “thin characterization of reflective judgment” (p. 7), “leaving thicker versions of the defense to the side for now” (p. 9). They do this in part because the thinly characterized judgments are “most easily tractable by means of experimental tools – the tools we intend to deploy” (p. 9). This is grist for the mill of critics who view philosophical case analysis as an extended, reflective, socially mediated activity and who object that X-Phi's experimental studies lack ecological validity.



### 4.3.3 *Expertise*

The above points lead directly to the related issue of philosophical expertise. The initial expertise-based objection to both positive and negative X-Phi ambitions was that, because the subjects of X-Phi surveys were undergraduates with little to no training in philosophy, the surveys were not informative about the epistemology and psychology of professional philosophical methodology. Critics motivated this objection with suggestive analogies. Ludwig (2007) observed that the discipline of mathematics is unthreatened by survey-studies revealing distorting influences on math amateurs. And Williamson (2011) pointed out that “we do not expect physicists to suspend their current projects in order to carry out psychological investigations of their capacity as laboratory experimentalists, on the basis of evidence that undergraduates untrained in physics are bad at conducting laboratory experiments” (p. 217).<sup>21</sup>

Pushing past these analogies, what is it that experts (specifically philosophy experts) have that novices lack such that empirically scrutinizing the novice’s use of cases will not inform about the pro’s use of cases? How one responds to this question will (or should) be informed by how one understands the purpose of philosophical case analysis in the first place – whether that method is used to delineate the borders of one’s own (or the folk’s) concepts (ala the Canberra Plan) or to illuminate the nature of objective structures and kinds (ala the “synthetic” aims described in Papineau (2009), Devitt (2012), and Kornblith (2002); the “material”, as opposed to “conceptual”, aims described in Machery (2017); see also Section 4.1 above). Moreover, we should not assume uniformity across philosophical sub-disciplines on this matter; some sub-disciplines (e.g., philosophy of biology, philosophy of psychology, naturalized ontology and

---

<sup>21</sup> See Hales (2006) and Sorenson (2014) for more analogies along these lines. See Nado (2015a) for discussion.

epistemology) likely espouse mostly synthetic aims, while other sub-disciplines (e.g., logic and perhaps metaethics) are more likely to espouse conceptual aims.

At any rate, proponents of the expertise objection characterize the philosophy expert-novice skill difference in various ways. We can use the three-level model of philosophical case analysis to express the core idea: The philosophy expert has learned to progress more skillfully and knowledgeably through the three levels of case judgment, using those levels (and the cases that prompt their cycling) to generate classificatory judgments and theoretical commitments that, assuming synthetic aims, more accurately capture explanatorily important kinds and structures relative to what novices generate (with or without similar case prompts). Indeed, a developmental progression like this is (nearly) the whole point of philosophical training.

It deserves emphasis that the progression includes both a procedural and evaluative (or content) component. Regarding the procedural component, Williamson claims that philosophers learn how “to apply general concepts to specific examples with careful attention to the relevant subtleties” (2007, p. 191) and to “decompose the task of thought-experimentation into consciously discernible sub-tasks” (2011, p. 224). Horvath (2010) claims that training makes philosophers “much more sensitive to potential ambiguities, unclarities or incoherencies” (p. 467). And Ludwig (2007) states that trained philosophers are skilled at “sorting out the various confusing factors that may be at work” (p. 149) in a complex philosophical case. Applied to our judgment model, these types of procedural skills are often employed in the context of Level 2 judgments (e.g., as a means of interrogating Level 1 judgments), and they inform much Level 3 discourse as well.<sup>22</sup>

---

<sup>22</sup> For example, the emerging understanding, facilitated by procedurally skilled public philosophical debate, that Jackson’s knowledge argument, whatever its merits, was not guilty of committing the intensional fallacy. See Stoljar (2017, p. 10).

The evaluative (or content) improvement, which is not disconnected from procedural expertise, concerns the ability to judge correctly the probative value of different arguments given for different classificatory judgments and thereafter adjusting one's credences accordingly. As we saw in Section 4.1, Deutsch and Cappelen emphasize that case analysis involves a competition between arguments rather than between intuitions. Philosophical training teaches you the historical and vetted arguments. It also teaches you how to evaluate arguments and how to construct new ones, where this involves, among other things, evaluating the explanatory profiles of competing classificatory schemes (and what counts as a good explanation in the first place). Indeed, the purpose of many thought experiments just is to lay bare a theory's explanatory commitments: the Swampman case illustrates that teleosemanticist's unificatory theory of content excludes the possibility of behaviorally sophisticated creatures that have accidental origins; Gettier cases illustrate that the JTB theory of knowledge includes lucked-into true beliefs; and so forth. It is then a matter of argument and the judicious weighing of explanatory merit as to whether these exclusions and inclusions indicate epistemic deficiency. As Jackson (2011) explains in the context of Gettier cases:

true justified belief isn't suited to the special role we give knowledge in epistemology.

We want knowledge to be a kind of gold standard in epistemology. The state we are in if and only if things have gone right, epistemically speaking. True justified belief had seemed to be such a state. Gettier cases taught us that it need not be. (p. 475)

We should not expect lay persons to understand what would count as the "gold standard in epistemology" (especially if we are assuming synthetic, material aims), as lay persons are not in

a good position to know what theoretical itch “knowledge” is scratching.<sup>23</sup> To know *that* generally requires training in both the methods and history of philosophy.<sup>24</sup> Sosa (2015) makes a related point:

Compare how we have been led to classify the tomato as a fruit, not a vegetable, and the whale as a mammal, not a fish, even if that departs from how many conceive of fruits, vegetables, mammals, and fish. It may of course turn out that philosophers, some of us, are led to a distinctive concept or conception of knowledge through our interest in knowledge itself, and through our search for an account of its nature. [...] The disagreement between philosophers and folk respondents on street corners might conceivably be analogous to the early disagreement among scientists and ordinary folk on what counts as a fruit, or as a fish. (p. 16–17)

Again, the idea here is that the philosophy expert has the requisite training and historical knowledge to judge more accurately when explanatory gain trumps departure from semantic defaults and snap-judgments.<sup>25</sup> This is neither chauvinism nor dogmatism if the explanatory

---

<sup>23</sup> When Deutsch and Williamson say that Gettier cases are *counterexamples* – that these cases reveal the *fact* that there is not knowledge – we might view them as claiming that there is no question about the deficient explanatory value of theories of knowledge that classify otherwise.

<sup>24</sup> See especially Millikan (2012) on the importance of extended philosophical training in the history of philosophy in this sense.

<sup>25</sup> There is an emerging trend of labeling this sort of proposal “conceptual engineering”. In the context of synthetic aims, which I suspect are the dominant (and correct) aims, I submit that the label “conceptual engineering” threatens to be either misleading or trivial, as just about every philosopher who has ever advanced a new theory about a part of the world when old theories were descriptively inadequate becomes a “conceptual engineer”. Whether the label is appropriate or interesting for explicitly norm-driven category revisions (e.g., in the sense of Haslanger 2012) or for those who are invested in the Canberra Plan is perhaps another matter.

gains are real and the descriptions of kinds or structures more accurate (just as it is not chauvinistic or dogmatic for expert biologists to privilege the whale-excluding fish concept despite folk push-back).<sup>26</sup>

#### 4.3.3.1 *Three objections to the expertise criticism, and responses to the objections*

As the expertise criticism has figured prominently in debates about the metaphilosophical relevance of X-Phi studies, it is worth sketching some objections as well as responses to the objections.

The first is that expert philosophers' acquired theories and schemas are "just as likely to contaminate as they are to clarify" (Alexander 2012, p. 95; Weinberg et al. 2010). In response, Williamson (2013, p. 473–474), Kornblith (2014, p. 165–166), and Devitt (2011, Section 5.5) have each pointed out that given the theory-ladenness of observation generally, one can raise this skeptical worry for any intellectual endeavor or discipline.<sup>27</sup>

The second objection we already discussed in Section 3.1. As reported there, several experimental philosophers have argued that the developmental conditions in philosophy are the same as the developmental conditions in domains in which experts are deluded about the quality of their expertise. As explained in Section 3.1, this argument overlooks important differences between philosophical epistemic activities and non-philosophical epistemic activities, thus undercutting the generalization about virtuous developmental conditions. The argument also

---

<sup>26</sup> This evaluative (or content) component of expert skill is played out – and developed through repeated iterations of – all three levels of our judgment model. It manifests at Level 1 judgment in so far as expert's snap judgments grow increasingly conditioned by acquired knowledge schemas. See, e.g., Kornblith (2007) and Devitt (2011).

<sup>27</sup> To better support the theory contamination objection to the expertise defense, the restrictionist might address head-on the sorts of explanatory considerations put forward for the theories themselves (e.g., contesting whether justified true beliefs fall short of the "gold standard" in epistemology; Section 4.3.3 above).

overlooks empirical evidence indicating that non-philosophical experts who engage in tasks that are relevantly similar to philosophical case analysis (simulation, relational retrieval, grasping the importance of rare events, the discovery of new categories) are led to superior performance in direct feedback-deficient domains (Bach 2021). Still, one hopes for a positive, empirically informed account of the development of philosophical content expertise. Such an account is needed to address more forcefully the concern that expert philosophers' distinctive content judgments are tracking indoctrination rather than gold-standard explanatory categories – a concern raised by Starmans and Friedman (2020, p. 22). For attempts at an empirically informed developmental account of philosophical content expertise, see Nolan (2015) and Bach (2019, Section 4.3).

The third objection is advanced on empirical grounds and consists of several recent X-Phi survey studies that suggest distorting influences on specifically professional philosophers' case judgments.<sup>28</sup> These findings should concern philosophers, but here are five observations that mitigate their significance for philosophical methodology.

The first we have already covered, which is that this new wave of empirical studies, like the prior wave, lack ecological validity. As Kauppinen (2018) explains, “the experiments attempt to study an aspect of philosophical expertise *in isolation from the theoretical and practical context in which such expertise is originally exercised*” (p. 50, emphasis in original), and thus the findings of those experiments do not undermine the assumption that philosophers' skills enable reliable case analysis.

---

<sup>28</sup> These include Hitchcock and Knobe (2009), Schulz, Cokely, and Feltz (2011), Schwitzgebel and Cushman (2012, 2015), Tobia, Buckwalter, and Stich (2013), as well as Wiegmann, Horvath, and Meyer (2020).

Second, one can argue that the studies assume an overly simplistic view of philosophical expertise. Recall the above distinction between procedural and evaluative (or content) philosophical expertise. Probably all philosophy PhDs possess considerable procedural expertise – they understand distinctions, focus on relevant features, recognize fallacies, and so on. However – and here we must remind ourselves of how difficult and unforgiving philosophy is – probably not all (maybe a minority) of professional philosophers are truly skilled at adjudicating between competing explanatory considerations.<sup>29</sup> While the new wave of X-Phi studies screens out those without procedural expertise, they likely fail to screen out those who lack content expertise. Analogously, consider that there are lots of sushi chefs these days (several at my local grocery, in fact), all with professional training. Consider also that high-end sushi work requires unusually sharp knives. In the hands of lesser sushi chefs, those knives are more likely to bring about nicked fingers and botched cuts than exemplary cuisine. The point is that the same item – a knife, a thought experiment – can function as an essential trade tool for one professional and a “disturbing” tool for another. But to register this point, one must first accept the quasi-elitist (if platitudinous) conception of philosophical expertise sketched above. Note that the resulting view is not incompatible with selective applications of restrictionism. For example, if we have reason to believe that a particular tool – a specific knife, a specific thought-experiment – is so poorly designed that it is likely to damage a high percentage of users, then perhaps that item should be

---

<sup>29</sup> That might sound elitist – and to be clear I don’t count myself as one of the skilled adjudicators – but it is also a sociological platitude. That is, most philosophers think that most other philosophers are wrong much of the time, and they think this not because they regard other philosophers as procedurally sloppy. See Bach (2019, Section 3.2) for discussion of the contrast between the distribution of content expertise and the distribution of procedural expertise.

taken off the market entirely (see also Dennett 1996). X-Phi studies will aid the discovery of such defective items.<sup>30</sup>

Third, X-Phi studies of professional philosophers have been mostly restricted to cases that have normative content (e.g., trolley and footbridge cases). There are independent reasons to be suspicious about philosopher's ability to think clearly about this domain (i.e., Mackie 1977), so it is a live possibility that these experimental findings will not generalize to professional philosophers' judgments about mind, epistemology, metaphysics, and other theoretical areas of philosophy (the objective targets of which may be more conducive to accurate philosophical theorizing). It also remains to be seen to whether the findings of this new wave of X-Phi studies can be replicated.

Fourth, there are X-Phi studies that appear to vindicate the objection based on expertise. For example, Starmans and Friedman (2020) show that both laypersons and non-philosophy academics are considerably more likely than academic philosophers to attribute knowledge in Gettier cases.<sup>31</sup> This is exactly what the expertise objection (particularly the synthetic version) predicts – that philosophy experts' developmental history has guided them to distinct (and more accurate) judgments about epistemological categories – what Jackson termed the “gold

---

<sup>30</sup> But given our inability to confirm the accuracy of first-order philosophical theoretical models (Section 3.1), we are strongly limited outside of the normal modes of (empirically-informed) philosophical argumentation in terms of what conclusive decisions we can make in this context (but see Levy 2013).

<sup>31</sup> Starmans and Friedman (2020) interpret the results of their study along these lines only for a “more focused” (p. 22) version of the expertise defense. They interpret their results as worrisome for what they consider to be the standard version of that defense (because, they say, the results show that non-philosophy academic experts judge differently than philosophers). As made clear both above and below, I think the best way (and the more common way) to formulate the expertise objection is the “focused” version. This version resources not only intellectual reflection (which appears to distinguish Starmans and Friedman's generic version) but also familiarity with philosophical explananda, established arguments, historical texts, and other components of a philosophical training regimen.



standard.”<sup>32</sup> This is analogous to the folk’s favoring of the whale-inclusive fish category versus the expert biologist’s grasp of which taxonomic schemes have greater explanatory payoff (a sure sign that those schemes are tracking the world’s real kinds and structures).<sup>33</sup>

Fifth, despite restrictionists’ claims to the contrary, we should not expect expert philosophers who (by hypothesis) use the method of cases reliably to be “insulated” from distorting influences. This point deserves separate discussion, below.

4.4 The reliable and expert-superior analysis of philosophical cases is compatible with that analysis being influenced by epistemically irrelevant factors

Several restrictionists claim that expertise-based defenses of philosophical case analysis require that experts’ judgments are sufficiently “inoculated,” “immune” to, or “insulated” from the influence of epistemically irrelevant factors. Various commentators on the expertise discussion have rejected this requirement.<sup>34</sup> Non-philosophical examples help demonstrate the general compatibility between reliability and instability. For example, Sorensen (2014) points out that “order effects have been demonstrated for master chess players and many other types of experts [...] Since the chess masters perform well despite these shortcomings, the impact on performance must be minor” (p. 137).

At issue here is whether X-Phi’s documented presentational and demographic effects are mere “foibles” (Ludwig 2018) – the sorts of general biases that you would expect to find accompanying any intellectual (or perceptual, see Sosa 2007) endeavor, including epistemically

---

<sup>32</sup> See Sosa (2015, p. 16–17); see also Papineau (2009), Jackson (2011), Strevens (2019), and Bach (2019).

<sup>33</sup> See Griffiths (1999) and Millikan (2000) on this point.

<sup>34</sup> See Sosa (2007), Devitt (2011), Williamson (2011), Papineau (2011), Nagel (2012), Kornblith (2014), Sorensen (2014), Rini (2015), and Kauppinen (2018).

fruitful and expert-superior endeavors – or whether they are symptomatic of an underlying epistemically deficient methodology.<sup>35</sup> At this dialectical juncture, restrictionists might ask for demonstration of the reliability (or “hopefulness”, see below) of the relevant philosophical methodology. However, if meeting that request presupposes confirmation of correct answers to first-order philosophical questions, then one might respond that the request illegitimately exploits the confirmational reality of philosophical analysis described earlier (Section 3.1).

#### 4.5 X-Phi restrictionist claims lead to over-generalization and self-defeat

Several philosophers have claimed that X-Phi’s negative program leads to an unacceptable and general form of skepticism about everyday and scientific judgment (Ludwig 2007, Sosa 2007, Williamson 2007, 2011, Horvath 2010). This critique is advanced in at least two ways. The first way claims that the inference from demographic and presentational effects to unreliability leads to an unacceptable skepticism. For example, if we view ordering effects as indicating epistemic unreliability, then it would seem that we need to view chess experts’ judgments about mating positions, which are also subject to ordering effects (Section 4.4), as unreliable. And if we treat demographic effects as indicating epistemic unreliability, then we seem forced to accept overly skeptical views about (at least) the status of scientific judgment. As Papineau (2011) observes in this spirit, “we wouldn’t expect physicists to throw up their hands in excitement just because somebody shows that different cultures have different views about the origin of the universe” (p. 83–84). The second way locates the generalizing factor at the level of the judgment-forming mechanisms. If the judgment mechanisms that drive philosophical case analysis are the same as

---

<sup>35</sup> Note the synergies between this and other criticisms of X-Phi; we should expect general psychological biases to manifest in contexts that diverge (e.g., are non-socially mediated, non-reflective, non-extended) from those in which evaluative philosophical expertise is standardly developed and employed.

the judgment mechanisms that drive the everyday application of general concepts, then deeming those mechanisms unreliable in the philosophical case invites an overly skeptical stance towards the everyday application of concepts.

In an extension, or application, of this overgeneralization critique, some philosophers charge that the restrictionist interpretation of X-Phi's experimental findings is self-defeating because that interpretation derives from philosophical concepts the justifications of which depend on the restricted evidential practices (see Williamson 2009, Horvath 2010, Ludwig 2018).<sup>36</sup> Because this charge is a species of the more general, over-generalizing criticism, we can focus on the latter.

The over-generalizing critique is directed at both intuition-based (Section 4.1) and non-intuition based (Section 4.2) forms of restrictionism. Weinberg (2007) is a notable and ambitious attempt to defend the former type of restrictionism from this objection. He claims that philosophers' evidential use of intuitions lacks the external corroboration and means of error detection needed to confer reliability to an evidential practice. On this view, the restrictionist challenge stays local to philosophers' uses of intuitions because scientific and perceptual evidential practices, unlike philosophers' evidential sourcing of intuitions, possess these epistemically "hopeful" features. It is fair to ask whether the epistemic standards that Weinberg imposes here capitalize on the confirmational predicament that applies to philosophical theory-building generally (Section 3.1, see also Sosa 2011). At any rate, readers should consult Horvath

---

<sup>36</sup> As Ludwig (2018) summarizes: "It [is] a point of mild irony that the use of survey results to argue against the use of thought experiments in philosophy relies on an epistemology which could only be supported by the sources which it aims to undermine" (p. 399).

(2010), Grundmann (2010), Brown (2011), Sosa (2011), Ichikawa (2012), and Nado (2015b) for discussion of whether Weinberg's attempt succeeds.<sup>37</sup>

In Section 4.1, I critically examined the descriptive claim that philosophers appeal to intuitions as evidence. In Section 4.2, following the arguments of Deutsch, Capellen, and Williamson, I suggested that a model of philosophical case judgment that did not psychologize evidence was more descriptively accurate. Machery (2017) advances such a model, claiming that “philosophical cases do not elicit attitudes distinct in kind from the application of concepts in everyday life” (p. 21). To escape the overgeneralization worry, this type of view needs to indicate why specifically philosophical cases interact with otherwise reliable judgment mechanisms to make them function unreliably. This is the point of Machery's discussion of the “disturbing characteristics” of philosophical cases (briefly mentioned in Section 4.2). Several commentators have claimed that this appeal to disturbing characteristics does not block the slide to skepticism. It is beyond the scope of this chapter to examine this interesting debate in detail, but readers can consult Machery (2011, 2017, Section 3.5), Williamson (2016), Deutsch (2020), and Nado (2020) for discussion.

Finally, it is important to observe that the overgeneralization objection has a distinct application to the developmental argument (see Sections 3.1 and 4.4 above) made by experimental philosophers against the reliability of expert philosophers' theories and theory-based judgments. Weinberg et al. (2010) claim that the epistemic virtue of expert philosophers' (theory-laden) intuitions cannot be defended in reference to the epistemic virtue of expert

---

<sup>37</sup> Note that the philosophical naturalist who construes a theory's explanatory and unificatory pay-off (and lack thereof) as an indicator of success (and error), and who views intuitions as both theory-laden and instrumental to the construction of further theories, has resources to claim that philosophers' evidential uses of intuitions are in this sense “hopeful”.

philosophers' theories because philosophers do not receive the good, direct kind of learning feedback that is needed for the development of epistemically virtuous theories.<sup>38</sup> Weinberg and colleagues (2010) are aware that this view suggests an overly skeptical stance towards philosophical theories generally. In response, they state that there are a few “successful” (p. 342) and “key” (p. 342) distinctions (they mention the use-mention and the epistemological-metaphysical distinctions) that “philosophers have wrung from their theorizing” (p. 342), that have “proved useful” (p. 351), and that are “exceptions that prove the rule” (p. 342).<sup>39</sup> It is not clear how this response is satisfactory; one can make similar claims about success and usefulness for just about any philosophical proposal (which is what philosophers do when they argue for their theories). What appears to distinguish Weinberg and colleagues' exempted theoretical distinctions is the extent of agreement among professional philosophers about the epistemic virtue of these distinctions. If that is what explains their privileged epistemic status, then it reframes Weinberg and colleagues' developmental argument against expert philosophical theorizing in a way that seems to confuse the conditions required for professional belief convergence, on the one hand, and the conditions needed for the development of accurate philosophical theoretical models, on the other. The defender of epistemically virtuous philosophical theories (and theory-laden judgments) does not need to make a case that the developmental conditions of philosophical discourse are conducive to professional agreement

---

<sup>38</sup> See Weinberg et al. (2010, p. 340–341).

<sup>39</sup> Weinberg and colleagues (2010) also state (p. 342) that we are very good at training philosophy students to use these distinctions correctly. That may be correct, but it is besides the point. One can also train students to be very effective and sensitive in their application of the rules of snake-oil medicine. What needs explaining is what accounts for the success – for example correspondence to natural kinds or properties – of the rules themselves, as well as what enabled expert philosophers to develop that theoretical success. See Bach (2019, p. 15–16) for further discussion.

(they are not). They only need to defend the claim that the developmental conditions are conducive to the construction of philosophical theoretical models (and associated classificatory judgments) that describe their targets with reasonable accuracy.

## **5. Conclusion**

Philosophy is difficult, which we already knew. But it is especially difficult if one is not taking the time to reflect carefully, if one is not engaging with epistemic peers in dialogue, if one is led into verbal disagreements, if one is less than highly skilled at evaluating the explanatory profiles of competing but empirically equivalent philosophical models, if one's judgments are expected to be insulated from general types of psychological bias, if one is expected to use psychological rather than objective data, or if one is not allowed use of conceptual practices considered trustworthy in everyday and scientific contexts. That, at least, is one way to understand an interdependent network of criticisms directed at X-Phi's (primarily) restrictionist evaluations of philosophical methods. According to these criticisms, X-Phi's evaluative claims, as well as the experimental simulations and descriptive models on which they are premised, impose additional epistemic burdens on philosophical case analysis, thereby construing an already challenging epistemic activity as considerably more challenging – and hence less trustworthy – than it in fact is.

Whatever we make of these criticisms, there is no question that the experimental studies and metaphilosophical explorations generated by and through experimental philosophy have improved our understanding of the psychology and epistemology of philosophical analysis, and moreover they inform how best to approach traditional philosophical questions. The extent of

this type of improvement, however, is ultimately limited by the empirical unverifiability of the accuracy of first-order philosophical theoretical models.

## References

- Alexander, Joshua (2010): “Is experimental philosophy philosophically significant?”. In: *Philosophical Psychology*, 23(3), 377-389.
- Alexander, Joshua (2012): *Experimental philosophy*. Cambridge: Polity Press.
- Alexander, Joshua & Weinberg, Jonathan M. (2007): “Analytic epistemology and experimental philosophy”. In: *Philosophy Compass* 2 (1):56–80.
- Bach, Theodore (2019): “In Defence of Armchair Expertise”. In: *Theoria*, 85(5), 350-382.
- Bach, Theodore (2021): “Why the Empirical Study of Non-philosophical Expertise Does not Undermine the Status of Philosophical Expertise”. In: *Erkenntnis*, 86(4), 999-1023.
- Bealer, George (1998): “Intuition and the Autonomy of Philosophy”. In Michael DePaul & William Ramsey (eds.), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Rowman & Littlefield. pp. 201-240.
- Bengson, John (2013): “Experimental attacks on intuitions and answers”. In: *Philosophy and Phenomenological Research*, 86(3), 495-532.
- BonJour, Laurence (1998): *In Defense of Pure Reason*. Cambridge, UK: Cambridge University Press.
- Boyd, Richard (2021): “Rethinking natural kinds, reference and truth: Towards more correspondence with reality, not less”. In: *Synthese*, 198(12), 2863-2903.
- Brown, Jessica (2013): “Intuitions, evidence and hopefulness”. In: *Synthese* 190 (12):2021-2046.
- Buckwalter, Wesley & Stich, Stephen (2013): “Gender and Philosophical Intuition”. In: *Experimental Philosophy, Vol.2*. Joshua Knobe & Shaun Nichols (eds.), Oxford University Press. pp. 307-346.
- Cappelen, Herman (2012): *Philosophy Without Intuitions*. Oxford University Press UK.
- Cappelen, Herman (2020): “Experimental philosophy without intuitions: an illustration of why it fails”. In: *Philosophical Studies* <https://doi.org/10.1007/s11098-020-01581-1>

Cappelen, Herman & Deutsch, Max Emil (2018): “The Crisis of Method in Contemporary Analytic Philosophy”. In: *Notre Dame Philosophical Reviews* 1 (05).

Clarke, Steve (2013): “Intuitions as Evidence, Philosophical Expertise and the Developmental Challenge”. In: *Philosophical Papers* 42 (2):175-207.

Colombo, Matteo, Georgi Duev, Michèle B. Nuijten, and Jan Sprenger (2018): “Statistical reporting inconsistencies in experimental philosophy”. In: *PloS one*, 13(4), e0194360.

Cova, Florian ; Strickland, Brent ; Abatista, Angela ; Allard, Aurélien ; Andow, James ; Attie, Mario; Beebe, James ; Berniūnas, Renatas ; Boudesseul, Jordane ; Colombo, Matteo ; Cushman, Fiery ; Diaz, Rodrigo ; van Dongen, Noah N'Djaye Nikolai ; Dranseika, Vilius ; Earp, Brian D. ; Torres, Antonio Gaitán ; Hannikainen, Ivar ; Hernández-Conde, José V. ; Hu, Wenjia ; Jaquet, François ; Khalifa, Kareem ; Kim, Hanna ; Kneer, Markus ; Knobe, Joshua ; Kurthy, Miklos ; Lantian, Anthony ; Liao, Shen-yi ; Machery, Edouard ; Moerenhout, Tania ; Mott, Christian ; Phelan, Mark ; Phillips, Jonathan ; Rambharose, Navin ; Reuter, Kevin ; Romero, Felipe ; Sousa, Paulo ; Sprenger, Jan ; Thalabard, Emile ; Tobia, Kevin ; Viciano, Hugo ; Wilkenfeld, Daniel & Zhou, Xiang (2018): “Estimating the Reproducibility of Experimental Philosophy”. In: *Review of Philosophy and Psychology* (1):1-36.

Cullen, Simon (2010): “Survey-Driven Romanticism”. In: *Review of Philosophy and Psychology* 1 (2):275-296.

Davis, Kathy (2008): “Intersectionality as buzzword: A sociology of science perspective on what makes a feminist theory successful”. In: *Feminist Theory* 9 (1):67-85.

Davis, Murray S. (1971): “That's interesting!: Towards a phenomenology of sociology and a sociology of phenomenology”. In: *Philosophy of the Social Sciences* 1 (2):309-344.

Davis, Murray S. (1986): “That's classic! The phenomenology and rhetoric of successful social theories”. In: *Philosophy of the Social Sciences* 16 (3):285-301.

Dennett, Daniel (1996): “Cow-sharks, Magnets, and Swampman”. In: *Mind and Language* 11 (1):76-77.

Deutsch, Max (2009): “Experimental philosophy and the theory of reference”. In: *Mind and Language* 24 (4):445-466.

Deutsch, Max (2015): *The myth of the intuitive*. Cambridge, MA: MIT Press

Deutsch, Max (2020): “The Method of Cases Unbound”. In: *Analysis*, Volume 80, Issue 4, Pages 758–771, <https://doi.org/10.1093/analys/anaa052>

Devitt, Michael (2012): “Whither Experimental Semantics?”. In: *Theoria : An International Journal for Theory, History and Foundations of Science* 27 (1):5-36.



- Gick, Mary L., and Keith J. Holyoak (1980): "Analogical problem solving". In: *Cognitive psychology* 12, no. 3: 306-355.
- Goldman, Alvin I. & Pust, Joel (1998): "Philosophical Theory and Intuitional Evidence". In: *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry*. Michael Depaul & William Ramsey (eds.). Rowman & Littlefield, pp. 179-200.
- Griffiths, Paul E. (1999): "Squaring the Circle: Natural Kinds with Historical Essences". In: *Species: New Interdisciplinary Essays*. Robert A. Wilson (ed.), MIT Press. pp. 209-228.
- Grundmann, Thomas (2010): "Some hope for intuitions: A reply to Weinberg". In: *Philosophical Psychology* 23 (4):481-509.
- Hales, Steven D. (2006): *Relativism and the Foundations of Philosophy*. MIT Press.
- Hannon, Michael (2018): "Intuitions, reflective judgments, and experimental philosophy". In: *Synthese* 195 (9):4147-4168.
- Haslanger, Sally (2012): *Resisting Reality: Social Construction and Social Critique*. Oxford University Press.
- Heck, Richard (2018): "Speaker's Reference, Semantic Reference, and Intuition". In: *Review of Philosophy and Psychology* 9 (2):251-269.
- Hitchcock, Christopher & Knobe, Joshua (2009): "Cause and Norm". In: *Journal of Philosophy* 106 (11):587-612.
- Horne, Zachary & Livengood, Jonathan (2017): "Ordering effects, updating effects, and the specter of global skepticism". In: *Synthese* 194 (4):1189-1218.
- Horvath, Joachim (2010): "How (not) to react to experimental philosophy". In: *Philosophical Psychology* 23 (4):447-480.
- Ichikawa, Jonathan (2012): "Experimentalist pressure against traditional methodology". In: *Philosophical Psychology* 25 (5):743 - 765.
- Jackson, Frank (1998): *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford University Press.
- Jackson, Frank (2011): "On Gettier Holdouts". In: *Mind and Language* 26 (4):468-481.
- Kahneman, Daniel, & Klein, Gary (2009): "Conditions for intuitive expertise: A failure to disagree". In: *American Psychologist*, 64(6), 515–526.
- Kauppinen, Antti (2007): "The rise and fall of experimental philosophy". In: *Philosophical Explorations* 10 (2):95 – 118.

Kauppinen, Antti (2018): "Who's Afraid of Trolleys?". In: *Methodology and Moral Philosophy* Jussi Suikkanen & Antti Kauppinen (eds.). Routledge, pp. 49-72.

Kelly, Thomas (2016): "Disagreement in Philosophy". In: *The Oxford handbook of philosophical methodology*. Cappelen, Herman, Gendler, Tamar, & Hawthorne, John P. (eds.). Oxford University Press, pp. 374-394.

Kneer, Markus, Colaço, David, Alexander, Joshua, & Machery, Edouard (forthcoming): "On second thought: Reflections on the reflection defense". In: *Oxford Studies in Experimental Philosophy, Volume 4*. Lombrozo, Tania, Nichols, Shaun, & Knobe, Joshua (eds.). Oxford University Press.

Knobe, Joshua (2003): "Intentional Action and Side Effects in Ordinary Language". In: *Analysis* 63(3):190-194.

Knobe, Joshua/Shawn Nichols (2017): "Experimental Philosophy". In: *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/>>.

Kornblith, Hilary (2002): *Knowledge and its Place in Nature*. Oxford University Press.

Kornblith, Hilary (2007): "Naturalism and intuitions". In: *Grazer Philosophische Studien*, 74(1).

Kornblith, Hilary (2014): *A naturalistic epistemology: Selected papers*. OUP Oxford.

Kuntz, Joe R., & Kuntz, J. R. C. (2011): "Surveying philosophers about philosophical intuition". In: *Review of Philosophy and Psychology*, 2(4), 643-665.

Levy, Neil (2013): "Intuitions and experimental philosophy: comfortable bedfellows". In: *Philosophical Methodology: The Armchair or the Laboratory?* Matthew C. Haug (ed.), Routledge, pp. 381-397.

Ludwig, Kirk (2007): "The Epistemology of Thought Experiments: First Person versus Third Person Approaches". In: *Midwest Studies in Philosophy* 31 (1):128-159.

Ludwig, Kirk (2010): "Intuitions and relativity". In: *Philosophical Psychology*, 23(4), 427-445.

Ludwig, Kirk (2018): "Thought Experiments in Experimental Philosophy". In: *The Routledge Companion to Thought Experiments*. Mike Stuart, James Robert Brown & Yiftach J. H. Fehige (eds.). New York: Routledge, pp. 385-405.

Lycan, William G. (2017, November): "On Evidence in Philosophy". In: *Proceedings and Addresses of the American Philosophical Association* (Vol. 91, pp. 102-117). American Philosophical Association.

- Machery, Edouard (2017): *Philosophy within its proper bounds*. Oxford University Press.
- Machery, Edouard (2011): “Thought Experiments and Philosophical Knowledge”. In: *Metaphilosophy*, 42: 191–214. DOI: 10.1111/j.1467-9973.2011.01700.x
- Machery, Edouard ; Mallon, Ron ; Nichols, Shaun & Stich, Stephen (2004): “Semantics, cross-cultural style”. In: *Cognition* 92 (3):1-12.
- Mackie, Jonh L. (1977): *Ethics: Inventing right and wrong*. Harmondsworth: Penguin.
- Millikan, Ruth G. (2000): *On clear and confused ideas: An essay about substance concepts*. Cambridge University Press.
- Millikan, Ruth (2012): “Accidents”. In: *Proceedings and Addresses of the American Philosophical Association*, 86(2), 92-103
- Nado, Jennifer (2015): “Philosophical expertise and scientific expertise”. In: *Philosophical Psychology* 28 (7):1026-1044.
- Nado, Jennifer (2015b): “Intuition, philosophical theorizing, and the threat of skepticism”. In *Experimental Philosophy, Rationalism, and Naturalism: Rethinking Philosophical Method*. Fischer, Eugen & Collins, John (eds.). Routledge, pp. 212-229.
- Nado, Jennifer (2016): “The intuition deniers”. In: *Philosophical Studies* 173 (3):781-800.
- Nado, Jennifer (2020): “Philosophizing out of bounds”. In: *Philosophical Studies*, 1-9.
- Nagel, Jennifer (2012): “Intuitions and Experiments: A Defense of the Case Method in Epistemology”. In: *Philosophy and Phenomenological Research* 85 (3):495-527.
- Nahmias, Eddy, Stephen G. Morris, Thomas Nadelhoffer, and Jason Turner (2006): “Is Incompatibilism Intuitive?”. In: *Philosophy and Phenomenological Research* 73 (1):28 - 53.
- Neisser, Ulric (1976): *Cognition and reality: Principles and implications of cognitive psychology*. W H Freeman/Times Books/ Henry Holt & Co.
- Nolan, Daniel (2015): “The a posteriori armchair”. In: *Australasian Journal of Philosophy*, 93(2), 211-231.
- Papineau, David (2009): “The poverty of analysis”. In: *Aristotelian Society Supplementary Volume* 83 (1):1-30.
- Papineau, David (2011): “What is X-Phi good for?”. In: *The Philosophers' Magazine*, (52), 83-88.

Paul, Laurie A. (2012): "Metaphysics as modeling: the handmaiden's tale". In: *Philosophical studies*, 160(1), 1-29.

Pust, Joel (2019): "Intuition". In: *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2019/entries/intuition/>.

Rose, David & Danks, David (2013): "In Defense of a Broad Conception of Experimental Philosophy". In: *Metaphilosophy* 44 (4):512-532.

Rini, Regina A. (2015): "How not to test for philosophical expertise". In: *Synthese* 192 (2):431-452.

Ryberg, Jesper (2013): "Moral intuitions and the expertise defence". In: *Analysis* 73 (1):3-9.

Schulz, Eric ; Cokely, Edward T. & Feltz, Adam (2011): "Persistent bias in expert judgments about free will and moral responsibility: A test of the Expertise Defense". In: *Consciousness and Cognition* 20 (4):1722-1731.

Schwitzgebel, Eric & Cushman, Fiery (2012): "Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers". In: *Mind and Language* 27 (2):135-153.

Schwitzgebel, Eric, & Cushman, Fiery (2015): "Philosophers' biased judgments persist despite training, expertise and reflection". In: *Cognition*, 141, 127-137.

Seyedsayamdost, Hamid (2015): "On gender and philosophical intuition: Failure of replication and other negative results". In: *Philosophical Psychology*, 28(5), 642-673.

Shanteau, James. (1992). Competence in experts: The role of task characteristics. *Organizational behavior and human decision processes*, 53(2), 252-266.

Sorensen, Roy A (2014): "Novice thought experiments". In: *Intuitions*. Robert Booth, Anthony & P. Rowbottom, Darrell (eds.). Oxford University Press UK, pp. 135-147.

Sosa, Ernest (2007): "Experimental philosophy and philosophical intuition". In: *Philosophical studies*, 132(1), 99-107.

Sosa, Ernest (2009): "A defense of the use of intuitions in philosophy". In: *Stich and His Critics*. Michael Bishop & Dominic Murphy (eds.), Blackwell, pp. 101-112.

Sosa, Ernest (2011): "Can there be a discipline of philosophy? And can it be founded on intuitions?". In: *Mind & Language*, 26(4), 453-467.

Sosa, Ernest (2015): "Philosophical Intuitions and Metaphysical Analysis". In: *Discipline filosofiche*. 25 (1):9-22.

Starmans, Christina & Friedman, Ori (2020): "Expert or Esoteric? Philosophers Attribute Knowledge Differently Than All Other Academics". In: *Cognitive Science* 44 (7), p.e12850

Stich, Stephen & Tobia, Kevin P. (2016): "Experimental Philosophy and the Philosophical Tradition". In: *Blackwell Companion to Experimental Philosophy*. Justin Sytsma & Wesley Buckwalter (eds.). Wiley-Blackwell, pp. 5-21.

Stoljar, Daniel (2017): *Philosophical progress: In defence of a reasonable optimism*. Oxford UK: Oxford University Press.

Strevens, Michael (2019): *Thinking off your feet: How empirical psychology vindicates armchair philosophy*. Harvard University Press.

Stuart, Michael T. ; Colaço, David & Machery, Edouard (2019): "P-curving x-phi: Does experimental philosophy have evidential value?". In: *Analysis* 79 (4):669-684.

Swain, Stacey ; Alexander, Joshua & Weinberg, Jonathan M. (2008): "The instability of philosophical intuitions: Running hot and cold on truetemp". In: *Philosophy and Phenomenological Research* 76 (1):138-155.

Sytsma, Justin & Livengood, Jonathan (2015): *The Theory and Practice of Experimental Philosophy*. Broadview Press.

Tobia, Kevin ; Buckwalter, Wesley & Stich, Stephen (2013): "Moral Intuitions: Are Philosophers Experts?". In: *Philosophical Psychology* 26 (5):629-638.

Uhlmann, Eric L., David A. Pizarro, David Tannenbaum, and Peter H. Ditto (2009): "The motivated use of moral principles". In: *Judgment and Decision making* 4, no. 6 (2009): 479-491.

Weinberg, Jonathan M. (2007): "How to challenge intuitions empirically without risking skepticism". In: *Midwest Studies in Philosophy* 31 (1):318-343.

Weinberg, Jonathan M. (2016): Going Positive by Going Negative: On Keeping X-Phi Relevant and Dangerous. In: *A Companion to Experimental Philosophy*. Sytsma, Justin, and Wesley Buckwalter, (eds). John Wiley & Sons, pp. 71-86.

Weinberg, Jonathan M. (2017): "What is negative experimental philosophy good for?." In: *The Cambridge companion to philosophical methodology*. D'Oro, Giuseppina, and Søren Overgaard, (eds.). Cambridge University Press, pp. 161-183.

Weinberg, Jonathan M. ; Nichols, Shaun & Stich, Stephen (2001): "Normativity and epistemic intuitions". In: *Philosophical Topics*, 29 (1-2):429-460.

Weinberg, Jonathan M. ; Gonnerman, Chad ; Buckner, Cameron & Alexander, Joshua (2010): "Are philosophers expert intuiters?". In: *Philosophical Psychology* 23 (3):331-355.

Weinberg, Jonathan M. ; Alexander, Joshua ; Gonnerman, Chad & Reuter, Shane (2012): "Restrictionism and Reflection: Challenge Deflected, or Simply Redirected?". In: *The Monist* 95 (2):200-222.

Horvath, Joachim, Karina Meyer, and Alex Wiegmann (2020): "Intuitive Expertise and Irrelevant Options." In: *Oxford Studies in Experimental Philosophy Volume 3*. Lombrozo, Tania ; Nichols, Shaun & Knobe, Joshua (eds.). Oxford University Press, pp. 275-310.

Williamson, Timothy (2007): *The philosophy of philosophy*. Oxford: Blackwell.

Williamson, Timothy (2009): "Replies to ichikawa, martin and weinberg". In: *Philosophical Studies*, 145(3), 465-476.

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

Williamson, Timothy (2013): "Review of Joshua Alexander, Experimental Philosophy". In: *Philosophy*, 88(3), 467-474.

Williamson, Timothy (2016): "Philosophical criticisms of experimental philosophy". In: *A Companion to Experimental Philosophy*. Sytsma, Justin, and Wesley Buckwalter, (eds). John Wiley & Sons, pp. 22-36.

Woolfolk, Robert L. (2013): "Experimental Philosophy: A Methodological Critique". In: *Metaphilosophy* 44 (1-2):79-87.

Young, Iris Marion (1994): "Gender as seriality: Thinking about women as a social collective". In: *Signs: journal of women in culture and society*, 19(3), 713-738.