Why the Empirical Study of Non-Philosophical Expertise Does Not Undermine the Status of Philosophical Expertise

Theodore Bach (tbach@bgsu.edu)

Abstract. In some domains (meteorology, live-stock judging, chess, etc.) experts perform better than novices, and in other domains (clinical psychiatry, long-term political forecasting, financial advising, etc.) experts do not generally perform better than novices. According to empirical studies of expert performance, this is because the former but not the latter domains make available to training practitioners a direct form of learning feedback. Several philosophers resource this empirical literature to cast doubt on the quality of philosophical expertise. They claim that philosophy is like the dubious domains in that it does not make available the good, direct kind of learning feedback, and thus there are empirical grounds for doubting the epistemic quality of philosophical expertise. I examine the empirical studies that are purportedly bad news for professional philosophers. On the basis of that examination, I provide three reasons why the empirical study of non-philosophical expertise does not undermine the status of philosophical expertise. First, the non-philosophical task-types from which the critics generalize are unrepresentative of relevant philosophical task-types. Second, empirical critiques of non-philosophical experts are often made relative to the performance of linear models – a comparison that is inapt in a philosophical context. Third, the critics fail to discuss findings from the empirical study of non-philosophical expertise that have more favorable implications for the epistemic status of philosophical expertise. In addition to discussing implications for philosophical expertise, this article makes progress in the philosophical analysis of the science of expertise and expert development.

1 Introduction

If we could empirically adjudicate between rival philosophical theories, then we would do so. But typically, the reason that we are engaging in philosophical theorizing and intuiting in the first place is that the relevant scientific experiments are unavailable (Sorenson, 1992; Paul, 2012).

There are, however, indirect ways of bringing empirical methods to bear on the epistemic status of philosophical theories and intuitions. One such method, popular with experimental philosophers, involves carrying out experiments on both undergraduate and professional philosophers that test whether epistemically irrelevant factors influence philosophical intuiting and theorizing. For the purpose of building a skeptical case against the value of philosophical expertise, this indirect method offers limited value. This is because interference effects are compatible with expert superiority (e.g., chess masters who are susceptible to ordering effects).

There is another and more forceful indirect method of challenging the epistemic status of expert philosophical intuiting and theorizing – the "Developmental Challenge" – that does not suffer this limitation. This method resources empirical studies of the performances of non-philosophical experts – performances for which success and failure *can* be directly empirically measured – and then claims that the developmental conditions of philosophy are very much like the developmental conditions in those domains in which the non-philosophical experts perform poorly.

In this essay, I will show that the empirical study of non-philosophical expert performance does not have the skeptical implication for philosophical expertise that the critics contend. In addition, I will make progress in the philosophical analysis of the science of expert development. In §2, I explicate the Developmental Challenge, focusing on the notion of direct feedback. In §3, I advance three distinct but mutually supporting critiques of the Developmental Challenge, and I explain why the empirical literature on non-philosophical expertise does not support a skeptical attitude toward the epistemic status of philosophical expertise. §4 concludes.

2 Explication of the Developmental Challenge

The Developmental Challenge (hereafter DC) is an argument that exerts dialectical pressure on philosophers who appeal to the enhanced epistemic quality of expert philosophers' intuitions and theories.[1] DC was first put forward in "Are Philosophers Expert Intuiters?" (2010), written by

[1] It is especially aimed at philosophers who use the "expertise defense" to diffuse experimental findings that show undergraduate philosophical intuitions as sensitive to epistemically irrelevant factors. For examples of the expertise defense, see Horvath (2010) and Williamson (2011).

Weinberg, Gonnerman, Buckner, and Alexander (WGBA). Steve Clarke, in his "Intuitions as Evidence, Philosophical Expertise and the Developmental Challenge" (2013), develops DC in greater detail. Ryberg (2013) also develops a version of DC.

WGBA's version of DC infers that there is no empirical support for the claim that philosophical training enhances the epistemic quality of philosophical theories and intuitive judgments. Absent such evidence, WGBA say that philosophers should stop invoking expertise to defend the epistemic quality of philosophical theories and intuitions. Clarke's version offers a stronger conclusion, inferring the inaccuracy of expert philosophical intuitions:

> Premise 1: In all (of a significant number of) examined domains where accurate professional intuitions have been acquired, clear, reliable and timely feedback is available to enable intuitions to be improved.
>
> Premise 2: Clear, reliable and timely feedback is unavailable to enable philosophers' intuitions to be improved.
>
> Conclusion: Therefore, it is very unlikely that professional philosophers have developed accurate professional intuitions.
> (Clarke, 2013, p. 192)

This article will focus on the above formulation of DC, though the arguments to follow apply equally to WGBA's and Ryberg's slightly different versions. Also, I will adopt the convention followed by empirical researchers and proponents of DC of defining an "expert" as someone who has experience and credentials in a given domain and who is regarded by peers as an expert.[2]

DC is an inductive argument. It generalizes from domains that have been empirically investigated with respect to particular characteristics – the quality of expert performance and related training conditions – to a domain that has not been empirically investigated with respect to these characteristics. Premise {1} describes a pair of correlations that emerge from the

[2] See Shanteau (1992) for elaboration.

empirical study of non-philosophical experts. The domains in which experts consistently outperform novices (e.g., meteorology, live-stock judging, chess) make available direct learning feedback to training practitioners, and the domains in which experts consistently fail to outperform novices (e.g., clinical psychiatry, long-term political forecasting, financial advising) do not make available direct learning feedback to training practitioners.

What is direct learning feedback? According to the critics, it is feedback that is "clear," "reliable," and "timely." It permits the "apprehension of repeated or salient successes and failures," and it allows "the expert-in-training to focus intently on practice, break the task down into components, and correctly diagnose (perhaps tacitly) the causes of success or failure" (Weinberg *et al*., 2010, p. 340). Summarizing the empirical literature, WGBA report that practitioners from the expert-credible domains are "confronted with a truly vast array of cases, with clear verdicts swiftly realized across a wide range of degrees of complexity or difficulty" (Weinberg *et al*., 2010, p. 341).

When discussing the nature of such feedback, the critics tend to focus on sensory or causal transactions with one's immediate environment:

> When a nurse has an intuition that a patient is in danger of an imminent heart attack … she can (and will) receive direct feedback from the environment. Either the patient does or does not go on to have an imminent heart attack. Learning whether or not her intuitions track reality enables her to train her future intuitions and improve these. (Clarke, 2013, p. 193)

Heart attacks are observable spatiotemporal phenomena with relatively clearly defined causal profiles. For this reason, they permit direct feedback for intuitions and theories with conceptual content pertaining to heart attacks.

Premise {2} claims that training philosophers are not sufficiently exposed to this epistemically good, direct kind of feedback. Contrasting the philosopher with the nurse, Clarke claims that:

> 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. (Clarke, 2013, p. 193)

Similarly, WGBA reflect on the various aspects of a philosophical training regimen and they find nothing that can play the direct-feedback role.

The conclusion of DC generalizes to the quality of philosophical expertise, claiming that it is likely suspect. It is likely suspect in the same way that expertise in clinical psychiatry and stock brokerage is suspect: institutional experts perform no better than institutional novices.

Both Clarke and WGBA claim that the deficient quality pertains to intuitive judgements. A reasonable way to think about intuitions (but not the only way of course) is as salient expressions of internalized theories.[3] WGBA (p. 344) grant this specific construal of intuitions. In fact, in the context of rebutting the expertise defense, WGBA advance DC specifically to handle this construal. WGBA claim that exposure to direct feedback is required to enhance the epistemic quality or accuracy of whichever theoretical knowledge structures underwrite the production of intuitive judgments.[4]

[3] Examples of this view include Kornblith (1998), Devitt (2012), Papineau (1996), and Kahneman and Klein (2009).

[4] For example, on the question of whether expert philosophers have developed "epistemically virtuous concepts or rules" (Weinberg *et al*., p. 341) – concepts or rules that they allow are candidates for underwriting epistemically virtuous intuitions – WGBA press the worry that philosophers "do not receive anything like the kind of substantial feedback required for such virtuous tuning" (*ibid*., p. 341). As to what counts as substantial feedback, WGBA refer back to the empirical literature on expertise: "It is important to keep the relevant contrast domains firmly in mind here, for they are what will provide the meterstick by which we can evaluate just what works as the right kind of feedback, and in what needed amounts, in order to produce effective training of real expertise. The fields in which competent experts routinely develop are those like meteorology, livestock judging, and chess" (*ibid*., p. 341).

The stakes for this debate are thus high: if DC is rationally persuasive, then we should not have confidence that expert philosophers are better positioned than novice philosophers to judge or possess epistemically superior theories.

Before moving to a critique of DC, I want to be clear about the scope of my discussion. In a series of articles, Weinberg and colleagues raise numerous concerns for the status of philosophical expertise and traditional philosophical methods. I will not be addressing here most of these arguments. For example, I will not attempt here a detailed account of the cognitive mechanisms that underwrite intuiting and theorizing,[5] and I will not be contesting recent surveys that reveal how professional philosophers are vulnerable to interference effects. In other words, I am not offering here a comprehensive defense of philosophical expertise. Nevertheless, DC arguments exhibit a trumping power in current debate over philosophical expertise (Clarke, 2013, pp. 188-192).[6] By evaluating DC then, I am addressing an issue that has considerable dialectical importance for current discussions about philosophical expertise and philosophical method.

I also want to be clear about the quality standard that is most relevant to an evaluation of DC. Proponents of DC often discuss the purported deficient quality of philosophical expertise in terms of the instability of experts' intuitive judgements (i.e., the sensitivity of intuitive

---

[5] See Weinberg (2007) for a discussion of this explanatory burden for defenders of philosophical intuition.

[6] Consider, for example, Devitt's assertion that a defense of philosophical expertise "requires only that the philosophers' intuitions be better, in general …, even if just as influenced by non-truth-tracking factors as the folk's" (Devitt, 2012, p. 22). Williamson (2011, pp. 218-219) and Sosa (2007) make similar claims. Even if this view is correct, it would not assuage the concern for philosophical expertise raised by DC. DC resources empirical data to establish a specific causal mechanism – direct or environmental feedback – for the development of enhanced expert performance, and then it claims that this mechanism is not present in the domain of philosophy. Any defense of philosophical expertise must confront this empirical challenge – it must show either that the causal mechanism is not needed for the development of virtuous philosophical expertise in the way that the critics contend or that the domain of philosophy does somehow make available direct feedback. This article develops the former type of response.

judgements to epistemically irrelevant factors). But a second and perhaps more fundamental standard – one that the first standard may or may not be a reliable indicator of – is the relative accuracy of experts' underlying philosophical theories and associated (e.g., theory-laden) intuitions. To see how these two standards can come apart, consider that chess masters who are primed with an unrelated prior problem will often abandon the correct solution to a current chess problem (Bilalić, McLeod, and Gobet, 2008; Saariluoma, 1990). However, no one should doubt whether chess masters possess more accurate theoretical models of chess strategy than chess novices. The arguments below will primarily focus on this second quality standard. That is, I will investigate whether the empirical literature on non-philosophical expertise provides compelling reasons for inferring that professional philosophers are unlikely to develop philosophical theories and intuitions that are more accurate than novice theories and intuitions.[7]

[7] This focus is a natural fit for defenders of philosophical expertise who already acknowledge distorting influences on experts' intuitive judgements (e.g., Williamson, Devitt, and Sosa; see fn. 6). There is also substantial overlap between this focus and how the critics frame the target of DC. Clarke's formulation of DC, provided in the quote above, is explicit that the target is whether expert philosophers' intuitions are accurate (Clarke, 2013, p. 192). Ryberg (2013, p. 4) is also clear about the importance of looking beyond distorting effects for the purpose of evaluating the reliability of philosophical expertise. In fact, Ryberg's version of DC – particularly its focus on what Ryberg terms "the quality assumption" – derives much of its force from the idea that trained philosophers do not know when their intuitions or theories meet this second standard. In Ryberg's terms: "While the philosopher may have engaged in many cases of intuition-based reasoning, it seems much less plausible to hold that she has prior experiences of having made intuitive judgements which led to *correct* moral answers" (Ryberg, 2013, p. 8). While WGBA often focus on the stability of intuitions, they invoke this second standard when discussing whether expert philosophers have developed "epistemically virtuous concepts or rules" (p. 341), philosophical theories that are "successful" (p. 342), and philosophical theories that are "key" (p. 342). Also relevant here is discussion from Alexander and Weinberg (2014) regarding different senses of the term "reliability" as used in debates over the epistemic status of philosophical intuition.

Finally, I submit that a philosophical analysis of the science of expert development – an analysis that identifies implicit methodological assumptions and improves our understanding of central theoretical concepts – offers epistemic value that is independent of its implications for debates about philosophical practice. An important contribution made by WGBA and Clarke, as I see it, is to shine a philosophical light on an important but neglected empirical literature. In addition to exploring the implications of this empirical literature for philosophical practice, I aim to make progress in the philosophical analysis of the science of expertise and expert development.

3 Why the Empirical Study of Non-Philosophical Expertise Does Not Undermine the Epistemic Quality of Philosophical Expertise

I will focus on three problems that make DC unpersuasive.

First, the inductive inference is too weak to warrant the argument's conclusion. The non-philosophical tasks for which direct feedback correlates with enhanced expert performance are quite specific. These tasks do not resemble most philosophical tasks as typically conceived. Thus, the non-philosophical task-types from which the argument generalizes are unrepresentative of relevant philosophical task-types.

Second, empirical critiques of non-philosophical experts are usually made relative to the performance of linear models. As I will explain, this comparison between human expert and computer algorithm is generally inapt in a philosophical context.

Third, there is less empirical support for premise {1} than the critics suggest. This is because, according to the same empirical literature referenced by the critics, there are various tasks for which the development of enhanced expert performance does not centrally depend on training against direct feedback. Moreover, the types of tasks for which reliable expert performance is less dependent on direct feedback share more relevant similarities with the types of theoretical tasks in which philosophers engage.

3.1 A Biased Sample

We can start with the claim that DC's inductive inference is too weak to warrant its conclusion. Clarke and WGBA focus on several studies and empirical overviews of expert performance – Camerer and Johnson (1991), Dawes (1994), Garb (1989), Kahneman and Klein (2009), and Shanteau (1992) – to substantiate their claim that the quality of expert performance depends on the directness of feedback (premise {1} of DC). It is true that these studies and reviews reveal a correlation between exposure to direct feedback and the quality of expert performance. But it is also true, and not discussed by the critics, that these studies and reviews generally advance this correlation in the context of specific task-types: predictive tasks, intervention-based tasks, and tasks that require classification into pre-set categories. When we look more closely at these task-types, two things become clear. First, these tasks have specific characteristics that make direct or environmental feedback valuable for effective training. Second, it is far from clear and arguably implausible that the theory-based tasks of philosophy share these same characteristics. I first describe each task-type and raise questions about whether it is similar or different in kind than relevant philosophical task-types. Then, I explain why these differences threaten DC's generalization about the developmental role of direct or environmental feedback.

Predictive tasks

Recall Clarke's contrast between intuiting nurses and intuiting epistemologists. Nurses profit from direct feedback "from the environment," allowing them to "directly discover" whether their intuitions are accurate. Epistemologists, on the other hand, "cannot directly discover" whether their intuitions are accurate, and this is because "direct feedback from the environment is typically unavailable." But what, precisely, is present in the first case that is absent in the second case? In the first case, the nurse is tasked to predict a future spatiotemporal event – a heart attack. Observing what does or does not in fact happen in the future, and then comparing that observation to one's past prediction, is what provides direct feedback about the quality of the prediction.

In fact, a great deal of the empirical literature emphasizing the developmental importance of direct or environmental feedback focuses specifically and often exclusively on experts who perform predictive tasks. This is the literature that inspires DC. For example, WGBA rely heavily on Camerer and Johnson (1991) to build their case that direct feedback explains the

development of epistemically virtuous expertise and that its absence in philosophy explains (or strongly suggests) the non-development of epistemically virtuous philosophical expertise. But it is crucial to note that Camerer and Johnson's analysis focuses specifically on predictive tasks. Discussing the connection between lack of direct feedback and poor expert performance, Camerer and Johnson stress that:

> Our arguments provide one possible explanation why knowledgeable experts, paradoxically, are *no better at making predictions* than novices and simple models. (Camerer and Johnson, 1991, p. 2010; emphasis added).

The bulk of the empirical literature on poor-performing experts shares this focus on predictive tasks. A small sample of other prominent examples would include Meehl (1954), which examines clinical predictions, Dawes (1971), which examines predictions about graduate students, Armstrong (1978), which examines economic forecasting, Carroll *et al*. (1982), which examines predictions about parole violation, and Tetlock (2005), which examines long-term political forecasting.

I do not challenge whether excelling at predictive tasks requires training in an environment that is rich in direct or environmental feedback. What I contest is that the relevant tasks of philosophy are predictive in this sense, or that philosophers' intuitions and theories are employed in the service of prediction. Consider philosophers who think that the goal of philosophical theorizing and intuiting is the production of general theories that *unify* scientific and/or conceptual theoretical frameworks. These philosophers will reject the idea that relevant philosophical tasks are predictive tasks, and they will also reject the notion that philosophical theories and intuitions have predictive aims. Similar claims apply for philosophers who are interested in modality, essence, conceptual analysis, and other non-predictive targets. Such philosophers should challenge the relevance of empirical generalizations drawn from a description of the developmental conditions required for the enhanced performance of predictive tasks.

Two examples will further explain the point. Philosophers who offer rival analyses of the concept of knowledge are not offering different material predictions. Rather, they are attempting to understand what it is that people are in possession of in those instances when they do have

knowledge (however frequent or infrequent).[8] Or consider philosophical theories about the nature of human motivation. Defenders of the theory of psychological egoism are not predicting that people will never jump on grenades, and proponents of rival theories, for example motivational pluralism, are not offering different predictions. Instead, proponents of these theories seek to explain how such behaviors, however often they occur, are structured with a variety of other psychological, biological, and evolutionary considerations. In other words, proponents of psychological egoism and motivational pluralism are not in the business of issuing specific behavioral predictions – they are in the business of advancing empirically adequate models of the nature of human motivation that favorably comport with various other epistemic considerations and bodies of knowledge.

To be clear, I am not claiming that philosophical theories can never be distinguished by their predictive content. For example, both Martí (2012) and Devitt (2015) describe how we might use data on actual linguistic usage to test philosophical theories of reference. And of course, some philosophers claim, contentiously, that at least some philosophical thought experiments are ways of testing in imagination the predictions of philosophical theories. But even here, we need to be cautious in how we understand the relationship between philosophical theory, associated intuitions, and material prediction. For example, the theory of psychological hedonism appears to predict that people would enter an "experience machine" if they believed that the machine would provide ideal pleasure. However, when people imaginatively simulate this opportunity, they generally report that they would not enter. Is this a failed material prediction of the theory of psychological hedonism? As others have already made clear, a variety of other causal factors – weakness of will, cognitive bias, the deterrent of painful deliberation – provide resources for explaining in hedonistic terms why people would not enter the machine. As Sober (2000) mentions in this context, rather than dictating specific behavioral outcomes philosophical theories like psychological hedonism indicate the *kind* of explanation to be given. Given such material flexibility, it is not surprising that debate in this area often centers on which theory best comports with selectionist principles (see, e.g., Sober and Wilson 1999 and the response in Stich 2007) – an observation that aligns with the unificatory focus of philosophical modeling as mentioned above.

[8] I further discuss the non-predictive aims of philosophical analyses of knowledge in §3.3.

Intervention tasks

Clarke's example about the nurse reminds us of an intervention-based task. Suppose that a paramedic has a theory or intuition that aspirin will forestall an immanent heart attack, and this directs the paramedic to intervene with aspirin. The paramedic then receives direct feedback about the success of this intervention in terms of whether the patient has a heart attack. Note the connection to prediction: one selects intervention *x* rather than intervention *y* on the basis of what one predicts would happen were one to *x* versus *y*. This type of intervention, whether it occurs during a chess match or in an emergency room, is one focus of empirical work on expert performance.

It should not surprise anyone that paramedics and other intervention professionals improve their intervention skills by intervening in the causal order or intervening within a game structured by a formal system and then observing what happens; the "directness" of the feedback fits the directness of the task. But again, we must ask: are most theory-driven projects of expert philosophers *intervention-based* in this sense? If they are not, then we should challenge attempts to restrict what can count as virtuous learning conditions for philosophy on the basis of what are virtuous learning conditions for intervention-based and formal system-based tasks. In other words, the same type of concern discussed above about generalizing from the development of virtuous predictive expertise to philosophical expertise applies to attempts to generalize from the development of virtuous interventionist and formal system-based expertise to philosophical expertise.[9]

[9] As was the case for predictive tasks, I do not mean to claim that philosophical theories and associated intuitions can never have intervention-based aims. For instance, there are philosophical models that recommend specific interventions toward contemporary social institutions (e.g., the institutions that control biomedical research – see Reiss and Kitcher 2009). But such intervention-based philosophical theories appear mostly restricted to particular philosophical sub-domains, for example the sub-domains of applied ethics and applied logics. While these domains do not appear to be the target of DC, perhaps they are the domains in reference to which DC is most cogent.

Classification tasks

Clarke relies on Dawes (1994) to build his version of DC. Dawes is concerned to show how clinical experience, which obviously expert clinical psychologists have much more of than novice clinical psychologists, does not often produce enhanced expert performance. The reason, claims Dawes, is that this professional experience does not provide direct feedback. However, closer examination of how Dawes understands the types of tasks for which direct feedback enhances performance reveals that these tasks, and the way that they function in skeptical arguments about expert performance, may have little to do with the activity and performance of expert philosophers.

According to Dawes (1994, p. 111), clinical work is best construed as a type of classificatory task, for example deciding whether to classify someone into the antecedently understood theoretical category "child abuser." Dawes further claims that this construal makes relevant empirical studies of people's classifying behavior. These studies, in turn, reveal the importance of direct feedback. Specifically, Dawes claims that laboratory experiments on people's ability to sort cards into gerrymandered categories reveals the importance of direct feedback.

For these experiments, the experimenter has a rule in mind, for example "blue triangles should be sorted to the left," and the experimental subject is tasked to develop sorting behavior that conforms to the rule. The experimental subject begins by randomly forming a hypothesis about the rule (Dawes, 1994, p. 113). Next, the subject sorts the cards according to that guess, making corrections to the guess on the basis of direct feedback ("correct sorting," "incorrect sorting") provided by the experimenter. If the subject does not receive this direct feedback, then he or she struggles to determine what rule the experimenter has in mind.[10]

[10] Readers might recognize these same laboratory experiments from Fodor (1975), where they figured centrally in his argument for an innate language of thought. More generally, Fodor resourced these experiments to argue that "learning" new semantic rules did not actually increase the richness of one's current conceptual system (see, e.g., Fodor, 1980, p. 148). Whether we

As was the case for intervention and predictive tasks, I do not challenge whether direct feedback is developmentally important for the classification of items into antecedently understood theoretical categories. What I contest is that the theory-based tasks of philosophy are akin to card-sorting tasks involving pre-set or antecedently understood (not to mention gerrymandered) categories.[11] Presumably, real or target philosophical success involves the formation and explanation of categories themselves as opposed to mere guess-work about how to sort items into pre-established and perhaps meaningless categories (I explore this idea further in §3.2 and §3.3). If philosophical tasks are not like card-sorting tasks, then we should challenge attempts to restrict what can count as virtuous learning conditions for philosophical tasks on the basis of what we know are virtuous learning conditions for card-sorting tasks.

So far, I have flagged three *prima facie* differences between the theory-driven projects of philosophy and the non-philosophical projects for which exposure to direct or environmental feedback is developmentally critical. There are, of course, many differences between two sets of tasks, most of which have zero relevance to DC's generalizing claim about the importance of direct or environmental feedback. For example, it does not undermine DC that philosophy differs from the domains investigated by expertise researchers with respect to which month of the year major conferences are held. Are there good reasons, then, to think that the three differences sketched above are relevant to, and thus undercut the grounding for, DC's generalizing claim about direct or environmental feedback?[12]

---

agree with Fodor about this, his view here serves as a warning that modeling philosophical activity in terms of sorting behavior could have radical or implausible implications.

[11] Perhaps there are some philosophical *training* tasks like this, for example having students sort cases relative to the use/mention distinction? Then again, there are astrological training tasks like this too, where students sort the ecliptic according to the twelve categories of the zodiac. The comparison helps make clear that the epistemic quality worth having pertains to the status of the categories themselves rather than one's ability to sort according to a rule.

[12] I thank a thoughtful anonymous reviewer for this journal for indicating the importance of this concern.

To see that there are such reasons, it is first helpful to observe a unity among the three differences. The card-sorting classificatory task just described is structurally similar to intervention-based tasks: the subject's guess leads to a sorting intervention which then invokes a negative or positive feedback response. As I noted earlier, intervention-based tasks are similar to predictive tasks: one's intervention is generally premised on a prediction or comparison of predictions. In fact, the three non-philosophical task-types considered above – prediction, intervention, and classification into pre-set categories – might usefully be framed as variations of a common, core task-type, namely: correctly predicting and/or rule-following in relation to a circumscribed system organized by causal regularities (e.g., weather systems), formal rules (e.g., mathematics, arbitrary classificatory rules), or some combination of these (e.g., competitive team sports). Put somewhat differently, these are tasks for which *success* is defined in terms of successful rule-following or prediction in reference to a specific and circumscribed causal or formal system. For convenience, I will call domains that centralize this type of success standard "causal/predictive domains."

When we step back and consider what makes an event or piece of information qualify as epistemically good feedback – that is, the type of feedback that causally explains the development of virtuous expertise – it is clear that it must carry accurate information to practitioners about the *successful* performance of whatever it is that they are supposed to do. This is what permits error correction and improvement over time. In causal/predictive domains, where success is generally a matter of predicting or manipulating a causal or formal system, it should not be surprising that epistemically good feedback has a "direct" or "environmental" component. Given what counts as success in these domains, there is nothing else for epistemically good feedback to be. But the situation is different in domains for which success is *not* understood in terms of predicting or manipulating some causal or formal system. For these "non-causal/non-predictive" domains, call them, epistemically good feedback would have an informational content that corresponds to the distinct success standards of these domains.

Now, a not uncommon view of philosophy is that its theory-driven projects have a goal or success standard that is fundamentally distinct from those that define causal/predictive domains. Paul (2012) is a clear statement of this view, describing how the subject matter and driving questions of most philosophical theories are ontologically prior to and distinct from those of scientific theories:

> Metaphysics is concerned to identify the real natures of the world while science is concerned to discover the range of *instances* of these natures….metaphysics gives the general and systematic story of what the categories are….The story from science is more specific … about how causing instances of different kinds and individuals involve different arrangements of physical properties and objects. (Paul, 2012, pp. 5-6)

So while the philosopher "looks to discover systematic, general truths," (*ibid*., p. 4) the more specific story from science is about "how to causally manipulate the world in order to bring [material] arrangements into existence" (*ibid*., p. 5). The contrast in subject matter between science and philosophy described by Paul here fits well with my earlier suggestions about the unificatory aims of philosophical models and the tendency for philosophical theories (e.g., theories of the nature of human motivation) to offer *kinds* of explanations rather than specific material explanations.

Putting this together, if epistemically good feedback is information about success, and if the goals of philosophical modeling are (generally) as described above, then it is a mistake to look toward the physical environment (or some token causal or formal system) for epistemically good philosophical feedback.[13] Where should we look instead? Given the *prima facie* systematic and unificatory purposes of philosophical modeling, perhaps a more promising place to locate indications of philosophical success is in the presence or absence explanatory coherence, unificatory power, fecundity, simplicity, and various other theoretical virtues that a philosophical model may or may not possess (see Nolan 2015 and Paul 2012 for recent defenses of this suggestion). This type of informational feedback would not be about the physical environment – it would be about relations among complex sets of representations, propositions, and theories –

[13] Which is not to say that the success of philosophical models is not constrained by what we know about the environment – philosophical models need to be empirically adequate. As Paul explains the point, "Science still acts as a constraint upon metaphysics – the metaphysician should want her theory of the whole world to be consistent with accepted scientific theories of the world – but it should not preemptively define the role or concepts of metaphysics. That would give us an understanding of reality that is exactly the wrong way around" (pp. 6-7).

and it would not be delivered directly through sensory observation. Thus, we should not hold its content and exemplification to those standards.[14]

But as mentioned in §2, my goal here stops short of providing a comprehensive defense of philosophical expertise. My goal is more modest – it is to clarify what can and cannot be inferred about the status of philosophical expertise from the empirical study of non-philosophical expertise. I thus leave the detailed arguments in support of the above metaphilosophical claims to others.[15] However, for dialectical purposes, it is important to point out that the same metaphilosophical explanatory burden applies to proponents of DC. If proponents of DC cannot show that philosophical theorizing *does* exemplify the same type of goal and success standard as theorizing in causal/predictive domains, then generalizing the distinctive success-based feedback standards of those domains to the philosophical domain would appear unjustified.[16]

---

[14] Which does not mean that the information is thereby magical or obscure. For example, Thagard's ECHO program (Thagard 1989), which detects various parameters of what Thagard terms "explanatory coherence," is an example of an attempt to make this general type of information computationally tractable. And as I explore in §3.3, there is some empirical support for the claim that experts improve performance by training against a non-direct or non-environmental type of informational feedback. (For those who might balk at the use of the term "information" here, it is worth flagging that philosophical accounts of natural information do not restrict information-carrying relations to causal relations – see, e.g., Dretske 1991).

[15] For more detailed discussion of the proposed contrast in subject matter between philosophy and science, see Paul (2012) and Lowe (2002). For a more detailed defense of the success-indicating role of theoretical virtues for philosophical modeling, see Nolan (2015), Paul (2012), and Swoyer (1999).

[16] Nor would it work simply to retreat to the common ground that predictive domains and philosophy share the same goal of tracking the truth. For one, this obscures the difference between the first-order and higher-order truths that informs the difference in subject matter suggested above. Second, it misses the distinction between amassing truths, on the one hand, and a successful theoretical organization of truths, on the other. A successful theoretical organization – one that separates truths that count as significant – is relative to the questions that guide the theoretical investigation (see especially Anderson 1995). As I have argued above, the questions

3.2 Finding the Right Comparison Class

No one thinks that an expert meteorologist who loses to a Laplacian demon in a weather predicting contest is for this reason a poor-performing expert. Laplacian demons, given their complete knowledge of the physical universe, are the wrong comparison class. For DC, the relevant comparison class for expert philosophers is novice philosophers. When the critics resource the empirical literature on expertise, they are challenging the default assumption that expert philosophers are better positioned than novice philosophers to judge or possess epistemically virtuous theories and intuitions.[17]

However, a major current that runs through the empirical literature resourced by the critics is based on a different comparison. It is based on the comparison between human experts and simple statistical, or regression, models. For example, Meehl's (1954) book *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence*, which has been influential in grounding skepticism about the quality of expert performance, reviewed studies comparing expert clinical predictions to the outputs of statistical models. This model versus human comparison is also central to the skeptical view of expertise developed in the heuristics and biases tradition (see Kahneman and Klein 2009). The upshot is that when empirical investigators of expertise claim that experts do not perform well, what they often mean by this is that the experts underperform relative to statistical models that are programmed to weigh, aggregate, and project particular variables.

The first thing to note here is that it is possible – indeed actual – for experts to outperform novices even if they underperform statistical models. To mention just one example, Johnson (1988) describes cases in which expert financial stock pickers – a domain that the critics describe

---

that guide philosophical investigations are often different in kind than those that guide non-philosophical investigations.

[17] See, e.g., Clarke: "Every professional philosopher started out as a non-philosopher, so defenders of this view are implicitly claiming that reliably accurate intuitions are acquired by a group of people who most probably started out with unreliable intuitions. We are owed an explanation of how this transformation happens" (Clarke, 2013, pp. 191-192).

as deficient in direct feedback – underperform statistical models while outperforming novices. The point is dialectically important: DC derives much of its empirical content from studies that are based on the comparison between human experts and statistical models, but these studies often leave open the possibility that the experts who underperform statistical models nonetheless outperform novices.

Second, and more worrying for DC I submit, is that the comparison between expert and statistical model has no plausible analogue in a philosophical context. In fact, *that* there is no analogue suggests something amiss about the strategy of generalizing from the empirical study of non-philosophical expertise to the epistemic status of philosophical expertise.

To appreciate this point, it is helpful to examine a case in which human experts have been shown to underperform statistical models. Dawes, Faust, and Meehl (1993) describe Pennsylvania's four-step procedure for determining whether to grant parole to offenders. The steps are: recommendations from correctional staff, recommendations from a parole case analyst, recommendations from a parole interviewer, and then ultimately the judgment of the parole board. At each layer, human experts scrutinize individual cases and employ subjective judgement with the goal of predicting criminal recidivism. Researchers were able to test these expert predictions by conducting follow up studies of rates of recidivism for offenders who were granted parole. The results were that the four-step expert procedure is quite poor at predicting recidivism. In fact, a simple three-variable regression model that considers an offender's number of previous convictions, offense type, and frequency of prison rule violation produces significantly more accurate predictions (and at a fraction of the cost). In the empirical literature on expert performance, this is a standard example of "experts who perform poorly."

But what enables the three-variable model to outperform the Pennsylvania Board of Probation and Parole? One reason is that its combining and weighing of variables is mechanically precise and free of interference from bias. But a second and more fundamental reason is that *the model is pre-packaged with the correct variables and correct explanatory relationship*. In other words, the model is valid, and its content tracks the features of the world that it is supposed to track. This means that a precondition of "experts performing poorly relative to statistical models" is that, at some prior point, someone had to figure out which algorithms and variables (e.g., previous convictions, offense type, and prison rule violation) are predictive of the target (e.g., recidivism).

This is not a controversial point. The following passage from Dawes – an empirical researcher who is generally critical of human expertise and whose research is referenced by proponents of DC – is quite clear about this essential contribution made by human experts:

> The statistical model may integrate the information in an optimal manner, but it is always the individual (judge, clinician, subjects) who chooses variables. Moreover, it is the human judge who knows the directional relationship between the predictor variables and the criterion of interest, or who can code the variables in such a way that they have clear directional relationships. And it is in precisely the situation where the predictor variables are good and where they have a conditionally monotone relationship with the criterion that proper linear models work well. The linear model cannot replace the expert in deciding such things as "what to look for," but it is precisely this knowledge of what to look for in reaching the decision that is the special expertise people have. (Dawes, 1979, p. 573)

In cases where we are already in possession of roughly the correct "theory" (e.g., about recidivism) as represented by the variables and algorithm of a valid statistical model, perhaps we are well advised to let computers rather than humans perform the computational labor. But if the goal is the correct delineation of theoretical variables and direction of explanation (more on this in section §3.3), then it is a mistake to overlook the essential epistemic contributions of human experts.

We can now see more clearly why the comparative evaluation of human experts and statistical models has no plausible analogue in a philosophical context. We do not require a detour into metaphilosophical abstraction to claim reasonably that much philosophical theorizing occurs at a level of analysis at which such prior determinations are not yet settled. That is, in much philosophical theorizing (whether about knowledge, causation, mental content, moral properties, etc.), we are *actively seeking out and debating the correct variables and combinatorial principles*. Put differently, the task is one of theory formation or selection rather than theory automation. And regarding theory formation and selection, empirical investigations of expert performance in non-philosophical domains have not shown that it is better performed

by statistical models. In fact, for reasons sketched above, the suggestion that it could be better performed by such models risks circularity.[18]

3.3 Not All Tasks Require Direct or Environmental Feedback for Enhanced Expert Performance

In §3.1 and §3.2 I argued that a difference between what constitutes success in causal/predictive domains and what constitutes success in philosophy – a difference reflected in the relative value of using statistical models in these domains – undercuts DC's generalization of the specific development requirements of causal/predictive domains to the philosophical domain. This section considers whether there is empirical information about the development of non-philosophical expertise that is more representative of, and would thus serve as a better basis for generalizing to, the development of philosophical expertise.

Given arguments put forward in §3.1 and §3.2, a more informative analogue for the development of philosophical expertise would be non-philosophical task-types the functions and success standards of which are similar to expert philosophers' theory-driven projects. This might include, for example, non-philosophical tasks that are pitched at a subject-matter level that is analogous to the prior, general, and systematic questions that I suggested are indicative of philosophical modeling. Or it might include non-philosophical tasks the performances of which would contribute instrumentally to the successful performance of philosophers' theory-driven projects. Such tasks would not centrally involve the manipulation or prediction of a token causal

[18] Related concerns about circularity (or at least explanatory breadth) are raised against researchers who widely apply probabilistic causal modeling. According to this concern, even if probabilistic causal modeling can explain how computational systems use probabilistic techniques to update hypotheses in light of new evidence, it fails to explain how computational systems generate these hypotheses in the first place (see, e.g., Goldman, 2006; Christie and Gentner, 2010). Chalmers, French, and Hofstadter (1992) discuss an even more general form of this concern in terms of the "hand-coded" representations often used by artificial intelligence researchers.

or formal system, and thus their enhanced performance (if that is the case) should not depend on receiving specifically direct or environmental feedback.

The empirical literature on non-philosophical expertise (including several sources referenced by the proponents of DC) explores several such non-philosophical task-types. I will discuss four of them: (1) knowing what to look for; (2) relational retrieval; (3) understanding the import of rare events and unusual factors; (4) simulating, imagining, and evaluating counterfactuals. Human experts demonstrate enhanced performance on these four tasks, and they do so even in domains that the researchers and proponents of DC flag as direct feedback-deficient. I briefly explain each task-type and its treatment in the empirical literature. I also indicate respects in which these task-types share more relevant similarities with theory-driven philosophical projects than do the direct feedback-dependent tasks described in §3.1.

Knowing what to look for

Recall that WGBA build their version of DC in reference to what they claim Camerer and Johnson (1991) report about the empirical study of expertise. But while Camerer and Johnson's article explains the importance of direct feedback for specifically predictive tasks (see the quote from §3.1), that article also states that experts who operate in direct feedback-deficient domains but who engage in *non-predictive* tasks can perform better than novices:

> The knowledge that experts acquire as they learn may not be useful for making better predictions about important long-range outcomes, but it may be useful for other purposes. Experts are indispensable for measuring variables (Sawyer 1966) and discovering new ones (E. Johnson, 1988). Furthermore, as experts learn, they may be able to make more kinds of predictions, even if they are no more accurate. (Camerer and Johnson, 1991, p. 210)

This idea that trained human experts have a unique skill-set for *discovering new variables* and making more *kinds of predictions* fits closely with the claims made at the end of §3.2. That is, while it may be the case that experts in direct feedback-deficient environments have trouble in projecting the values for certain categories, there are good grounds for claiming that these same

experts do well at knowing which categories are relevant in the first place and what are their causal-explanatory relationships.

It is worth returning to Dawes on this point, who in the following passage explores how a case study from Einhorn (1972) demonstrates the enhanced ability of human experts to discover explanatorily relevant variables:

> The distinction between knowing what to look for and the ability to integrate information is perhaps best illustrated in a study by Einhorn (1972). Expert doctors coded biopsies of patients with Hodgkin's disease and then made an overall rating of the severity of the process. The overall rating did not predict survival time of the 193 patients, all of whom died. (The correlations of rating with survival time were all virtually 0, some in the wrong direction.) The variables that the doctors coded did, however, predict survival time when they were used in a multiple regression model. In summary, proper linear models work for a very simple reason. People are good at picking out the right predictor variables and at coding them in such a way that they have a conditionally monotone relationship with the criterion. (Dawes, 1979, pp. 573-574)

Thus, while these medical experts who were working in a direct feedback-deficient context predicted poorly, the evidence for this comes from regression models the superior performance of which *presupposes* the enhanced performance of these same human experts on a distinct task. Specifically, these experts demonstrated enhanced performance with respect to locating correct causal/explanatory factors and understanding their direction of explanation.

This task – knowing what to look for and determining which factors are relevant in the first place for modeling a target phenomenon – looks an awful lot like what philosophers aim to do when advancing philosophical models. For example, should principles of quantum theory be brought to bear on questions about human free will? Should gender categories be understood as explicable on the basis of social structures rather than or in addition to physiological or sexual categories? Are debates over the status of folk-psychology beholden to recent advances in neurobiology? Getting clear on these types of questions would appear central to the job description of professional philosophers. Certainly these questions speak to the unificatory aims of philosophical theories as suggested in §3.1. If this is correct, then it is encouraging to the

professional philosopher that experts in direct feedback-deficient contexts who pursue analogous questions about explanatory relevance and theory formation demonstrate enhanced performance, even if these experts are later outperformed by linear models on the "applied" aspects of such tasks.[19]

Relational Retrieval

The importance of determining correct relevance relations highlights another component of the empirical literature on expert performance that is potentially supportive of professional philosophers. When novices and experts interpret a case or work on a problem, there is any number of representational schemas that they might activate and use toward an interpretation or solution. Empirical research on analogy and knowledge transfer has made considerable progress in understanding the cognitive processes that underlie this activity. That research reveals an important difference between experts and novices: experts are likely to transfer information (often automatically and unreflectively) on the basis of relational similarity between target and

[19] At least some intuitions appear to play a role in delivering judgements about explanatory responsibilities in this sense, thus involving (depending on how reflective the intuitions are) the cognitive unconscious to this end. For example, do theories of mental content have the explanatory responsibility of accounting for, in terms of contentful mental states, the *prima facie* intelligent behavior of a conscious creature that was just created in a swamp lightning storm? Those who "have the intuition" that so-called Swamp-persons are a problem for the theory of teleosemantics appear to judge that theories of mental content do have this explanatory responsibility (a responsibility that teleosemantics would fail). Those who do not "share the intuition" appear to reject that theories of mental content possess this explanatory responsibility. The difference in intuition here is not about what would happen, and it is generally agreed that teleosemantics must classify Swamp-persons as not having contentful mental states (this classification resulting from the simulative component of the thought experiment, as I mention below). The difference is in whether this classification should count as a theoretical vice – about whether Swamp-persons are a proper explanandum for philosophical theories of mental content. See Neander (1996) and Millikan (1996) for a related discussion along these lines.

stored knowledge, whereas novices are likely to transfer information (often automatically and unreflectively) on the basis of attributional or object-level similarity between target and stored knowledge.[20]

Researchers interpret novices' tendency for attribution-based retrieval as an epistemic failure, e.g., "the failure of relational retrieval" (Gentner *et al*., 2009, p. 1375), and they interpret experts' tendency for relational-retrieval as an epistemic strength.[21] This is for good reason. Across all domains, good explanations and accurate problem solving generally depend on sensitivity to the target's deep and systematic relational features rather than a fixation on its isolated or superficial features.[22] It would be surprising if this were not also true of philosophical explananda, especially if we assume the type of philosophical subject matter discussed in §3.1.

---

[20] Empirical support for this claim can be found in Novick (1988), Clement (1982, 1986), Chi, Feltovich, and Glaser (1981), Catrambone and Holyoak (1989), Lowenstein, Thompson, and Gentner (1999), Blanchette and Dunbar (2001), and Shafto and Coley (2003). See Gentner (1983) for a codification of the representational differences between attributes, relations, higher-order relations, and systematicity among relations.

[21] More generally, Gentner and colleagues characterize children's cognitive development in terms of their increasing ability to conceptualize domains through relational categories rather than object categories (i.e., "the relational shift"). In fact, Gentner (2003) argues that the capacity to develop relational knowledge through relational abstraction (coupled with a symbol system that can express and help develop that relational knowledge) is what distinguishes human cognition.

[22] Consider two math students who each attempt to solve a conditional probability problem about the sale of crude oil. One student is reminded of a previous problem that is structurally similar (it is about the probability of soil erosion under certain conditions), and transfers information (perhaps unconsciously) from a knowledge schema encoding that problem to the target problem. The other student is reminded of, and transfers knowledge from (perhaps unconsciously), a previous problem that is attributionally but not structurally similar (the problem involved the quadratic equation and the density of crude oil). The former student has employed the more virtuous problem-solving strategy.

In order to explain the relationship between expertise and relational retrieval, cognitive scientists appeal to experts' more *uniform relational encoding* of a domain (Gick and Holyoak, 1983; Forbus, Gentner, and Law, 1995; Lowenstein, Thompson, and Gentner, 1999; Gentner, 2003). Non-uniform relational encodings embed relational knowledge in particular exemplars and contexts.[23] Given their lack of domain experience and domain knowledge, novices tend to have idiosyncratic, object-specific representations rather than flexible relational representations. This type of encoding impedes the spontaneous retrieval of useful knowledge structures because it is difficult to match structural elements in the target problem with structural elements that are contextually embedded in potential source representations (Gentner, 2003; Bach, 2014).

But how is it that domain experience develops relationally explicit representations and thus more uniform relational encodings? A primary appeal of Gentner and colleagues' widely-accepted S*tructure-Mapping Theory* (SMT) is that it provides a rich, cohesive, and empirically supported set of answers to this question.[24] According to SMT, a primary means by which we acquire more uniform relational encodings, and thus greater ability for relational retrieval and transfer, is through the relational focus and relational abstractions that occur as a result of aligning mental representations. Representational alignments occur during the activity of *comparison*, for example when one compares the solar system and the Rutherford atom or, more mundane, two solar systems. SMT provides experimental evidence and computational models that indicate how learners have a tacit preference for relational structure when comparing items. This relational focus promotes relational abstraction and leads to more uniform relational encodings.

---

[23] For example, young children do not uniformly encode the relation of *unclehood* and instead conflate this relation with a particular uncle or men of a certain age. In contrast, someone with a uniform relational encoding of unclehood possesses a portable mental predicate that can be matched across superficially different object participants. Possession of this predicate would enable one to grasp that a certain newborn baby, as well as pipe-smoking Uncle Bill, are both *uncles*.

[24] SMT is a rich and complex theory developed over the last thirty years and through many articles and empirical studies. The discussions to follow are a simplification, but see Gentner and Colhoun (2010) and Bach (2012) for general overviews.

Importantly, and as demonstrated in empirical studies, the comparison-based relational abstractions that facilitate experts' increased tendency for relational retrieval occur in the absence of feedback (see especially Kotovsky and Gentner, 1996).[25] Relational abstraction is dependent, however, on opportunities to compare. According to Gentner and colleagues, a primary means through which learners are "invited" to engage in comparison-based learning is relational language (Gentner, 2003). And of course, philosophical discourse is rich in relational language.[26]

Clearly, much more needs to be said about the type of epistemic contribution that comparison-based abstraction and relational retrieval can make for training and expert philosophers. (In particular, a preference for relational retrieval would not be an epistemic advantage if the transferred relational claims were false). Nonetheless, it is important to observe that, given its subject matter and goals, there is considerable cross-fertilization in philosophical theory construction and evaluation. For example, it is not unusual for philosophical models (and critiques) of normative phenomena to import ingredients from the philosophy of biology, metaphysics, philosophy of mind, the history of philosophy, and non-philosophical domains. These types of epistemic transfers – as well as the use of thought experiments to invoke the unconscious transfer of representational schema (Gendler, 2007) – are *prima facie* more likely

[25] The abstractions are the outputs of a domain-general cognitive mechanism, computationally modeled by Gentner and colleagues' "Structure-Mapping Engine," that begins with a few processing constraints and is driven by invitations to compare.

[26] Philosophical problems might be understood through relational concepts and terms (e.g., "etiological function", "representation", "projectable predicate", "counterfactual dependence", "modus tollens", "reliable mechanism") or primarily through surface features, idiosyncratic object concepts, and context-based examples (e.g., a four chambered heart, the belief that tomatoes are red, green emeralds). As Gentner explains, "habitual use of a stable system of relational language can increase the probability of relational reminding. In instructional situations, it can foster appropriate principle-based reminding and transfer, and mitigate the perennial bugaboos of retrieval: inert knowledge and surface-based retrieval. The growth of technical vocabulary in experts reflects the utility of possessing a uniform relational vocabulary" (Gentner, 2003, p. 209).

successful if they are driven by relational retrieval rather than attribution-based retrieval. It is thus encouraging for professional philosophers that empirical research on non-philosophical expertise reveals how domain experience, even in the absence of feedback, confers an increased tendency for relational retrieval.

Understanding the import of rare events and unusual factors

Johnson (1988) discusses empirical investigations of expert securities analysts and stock pickers, which is a domain that the critics agree is deficient in direct feedback. Johnson reports that expert analysts generally performed better than novices, especially when given access to recent news items. Why is this? According to Johnson, the data indicate that the experts but not the novices are particularly good at judging the relevance and impact of rare cases, for example how the death of a CEO might impact the future price of a stock. The expert is better at grasping the meaning of the event, reasoning about its causal role, and then matching it to a broader pattern. In the context of predictive tasks, this focus on the particular tends to facilitate base-rate neglect, reflecting an epistemic weakness relative to statistical models. But in the context of many *non*-predictive tasks, where there is less concern about base-rate neglect, this heightened focus on the relationship between rare events and general patterns often leads to improved performance.

In the case of philosophical tasks that involve conceptual analysis and modal questions, there are especially good reasons for thinking that someone who is good at understanding the import of rare cases will have an epistemic advantage over someone who is less capable of reasoning carefully about such cases. In other words, just as expert stock-pickers, given their focus on and handling of rare cases, performed better than novices in a direct feedback-deficient environment, so too can expert philosophers, given their focus on and handling of rare cases, perform better than novices in the direct feedback-deficient domain of philosophy.

Consider Gettier's counterexample to the JTB theory of knowledge. If the goal of Gettier's discussion had been to *predict the frequency at which people form justified true beliefs about the number of coins in someone's pocket and yet lack knowledge*, then neither Gettier nor his readers were well served by the intense focus on the case of Smith and Jones. This is because the case of Smith and Jones is exceedingly rare at best, so a heightened focus on it would promote base-rate neglect and less accurate predictions than a linear model programmed with the

same predictive goal. Of course, such material predictions were never the goal of Gettier's discussion or the debate inspired by that discussion. Given the actual, conceptual goals of Gettier's discussion, he probably did quite well to focus on the rare event of Smith and Jones. In fact, whenever the target of philosophical analysis involves modal questions (e.g., about necessary conditions), it is likely important to have an enhanced ability for grappling with the import of rare cases. It is thus good news for expert philosophers that experts in non-philosophical domains, even if those domains are direct feedback-deficient, have been shown to be superior to novices in their ability to process the import of rare cases.[27]

Simulating, imagining, and evaluating counterfactuals

In order to develop and evaluate philosophical theories accurately, it is important to identify their theoretical implications and commitments. Thought-experimentation and counterfactual reasoning may play important roles here, teasing out a model's commitments so that they can be better evaluated. The earlier discussions of the "experience machine" and "swamp-persons" (fn. 19) showed how thought experiments might be used this way. If this is correct, then empirical research on non-philosophical experts' thought-experimentation and counterfactual reasoning can provide clues as to whether expert philosophers are likely more effective than novice philosophers when employing these same cognitive tools.

Nersessian (1993) and Gendler (1998, 2004) have explained how certain forms of thought-experimenting involve a type of mental modeling that relies on "constructing and making inferences from a mental simulation" (Nersessian, 1993, p. 292). Williamson (2007) has also argued that thought experiments function as a type of argument that recruits mental simulation for their evaluation. If these and related proposals accurately describe the cognitive mechanics involved in thought-experimentation and counterfactual reasoning, then the science of non-philosophical expertise as it pertains specifically to experts' use of mental simulation is

[27] It is worth noting that this analysis provides a more optimistic view of how the "unusual nature of philosophical cases" (Machery, 2017, p. 113) influences philosophical judgement than the view defended in Machery (2017).

relevant to an evaluation of expert philosophers' use of thought-experimentation and counterfactual reasoning.

Several lines of empirical evidence converge on the finding that non-philosophical experts, including those who work in direct feedback-deficient environments, more successfully employ mental simulation than do novices. Kahneman and Klein (2009) and especially Klein (1998) report on research that indicates how experts rely on mental simulation to perform more reliably during task performance. For example, in his investigation of expert fireground commanders, Klein reports that expert commanders were uniquely able to use simulation to produce new and accurate information toward enhanced task performance, for example, determining the best route into a burning building or how best to rescue someone trapped in a car.[28] Hogarth (2001) reports on how expert practitioners in direct feedback-deficient environments can hone their intuitions by using their imaginations to evaluate counterfactual scenarios.[29] These lines of empirical research are potentially supportive of expert philosophers who in certain contexts may rely on mental simulation and counterfactual reasoning to draw out the theoretical implications and commitments of candidate philosophical models.

4 Conclusion

This article defended the epistemic status of philosophical expertise against arguments derived from the empirical study of non-philosophical expertise. That defense was developed in three stages. I first argued that there is neither empirical nor theoretical support for the claim that the task-types described in empirical studies indicating the importance of direct or environmental feedback share relevant similarities with philosophical tasks. Second, I explained how the empirical case for underperforming non-philosophical experts is often made relative to statistical models, and I showed why this comparison is inapt in a philosophical context. Third, I explored examples from the empirical literature of non-philosophical experts who perform well even if they developed their expertise in direct feedback-deficient environments. I indicated how the

[28] See Klein (1998, pp 18-21; Ch. 5).

[29] Hogarth (2001, pp. 225-226).

types of tasks on which these experts performed well share relevant similarities with the investigative and theory-driven projects of expert philosophers.

I suggest that each stage of this defense is sufficient on its own to defend the epistemic quality of philosophical expertise against extant arguments derived from the empirical literature on non-philosophical expertise. But I also submit that the unity and relations of mutual support among the three defenses provides a particularly strong rebuttal to the skeptical case against expert philosophy that is based on the empirical study of non-philosophical experts.

At no point in making these arguments did I advance or rely on the claim that enhanced expert performance on philosophical tasks can develop in the complete absence of epistemic feedback. Rather, I argued that the science of non-philosophical expertise does not provide sufficient grounds for claiming that a particular type of feedback – so-called direct or environmental feedback that is produced by spatiotemporal investigative targets or in the context of formal systems – is a developmental requirement for epistemically virtuous philosophical expertise. It is thus consistent with my arguments to claim that alternative sources of epistemic feedback – for example a philosophical theory's explanatory and unificatory power (or lack thereof) – can provide information about the epistemic quality of underlying philosophical theorizing and intuiting, even if such forms of feedback would not qualify as so-called direct feedback. I leave the careful exploration of this and related positive proposals to another occasion.

I cannot fully defend here the claim that expert philosophers are better theorizers and intuiters than novice philosophers. But I can claim, and I hope to have shown in the preceding discussion, that we should not infer from the empirical study of non-philosophical expertise that expert philosophers perform no better than novice philosophers with respect to philosophical theorizing and intuiting.

Finally, by identifying implicit assumptions and clarifying concepts that are important for the evaluation of cross-disciplinary claims about the development of expert performance, I hope to have made progress in the philosophical analysis of the science of expertise.

Acknowledgements: I thank two anonymous reviewers from this journal for their helpful comments and suggestions. I also thank audience members at the 2018 Central Divisional

Meeting of the American Philosophical Association for feedback on research related to this article. Support for work on this article was generously provided by Bowling Green State University and the Office of the Provost in the form of Faculty Improvement Leave for Spring 2019.

References

Alexander, J., & Weinberg, J. M. (2014). The 'unreliability' of epistemic intuitions. *Current controversies in experimental philosophy*, 128-145.

Anderson, E. (1995). Knowledge, human interests, and objectivity in feminist epistemology. *Philosophical Topics*, *23*(2), 27-58.

Armstrong, J. S. (1978). *Long range forecasting: From crystal ball to computer*. New York: Wiley.

Bach, T. (2012). Analogical cognition: Applications in epistemology and the philosophy of mind and language. *Philosophy Compass*, *7*(5), 348-360.

Bach, T. (2014). A Unified Account of General Learning Mechanisms and Theory-of-Mind Development. *Mind & Language*, *29*(3), 351-381.

Bilalić, M., McLeod, P., & Gobet, F. (2008). Inflexibility of experts—Reality or myth? Quantifying the Einstellung effect in chess masters. *Cognitive psychology*, 56(2), 73-102.

Blanchette, I., & Dunbar, K. (2001). Analogy use in naturalistic settings: The influence of audience, emotion, and goals. *Memory & Cognition*, *29*(5), 730-735.

Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In K. A. Ericsson & J. Smith (Eds.),

Toward a general theory of expertise: Prospects and limits (pp. 195–217). Cambridge: Cambridge University Press.

Carroll, J. S., Winer, R. L., Coates, D., Galegher, J., & Alibrio, J. J. (1982). Evaluation, diagnosis, and prediction in parole decision making. *Law and Society Review*, 17, 199-228.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 1147–1156.

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence, 4*(3), 185-211.

Chi, M., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, *11*(3), 356-373.

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

Clement, J. (1982). Analogical reasoning patterns in expert problem solving. *Proceedings of the fourth meeting of the Cognitive Science Society* (pp. 79-81), Ann Arbor, MI.

Clement, J. (1986). Methods for evaluating the validity of hypothesized analogies. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 223-234), Amherst, MA. Hillsdale, NJ: Erlbaum.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist*, 26, 180-188.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34 (7), 57-582.

Dawes, R. M. (1994). *House of Cards: Psychology and Psychotherapy Built Upon Myth.* New York, The Free Press.

Dawes, R. M., Faust, D., & Meehl, P. E. (1993). Statistical prediction versus clinical prediction: Improving what works. *A handbook for data analysis in the behavioral sciences: Methodological issues*, 351-367.

Devitt, Michael. (2012). Whither experimental semantics? *Theoria* 72: 5-36.

Devitt, Michael. (2015). Testing theories of reference. *Advances in experimental philosophy of language*, 31-63.

Dretske, F. I. (1991). *Explaining behavior: Reasons in a world of causes*. MIT press.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational behavior and human performance*, *7*(1), 86-106.

Falkenhainer, B., Forbus, K.D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence, 41*, 1-63.

Fodor, J. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

Fodor, J. (1980). Fixation of belief and concept acquisition. In Piatelli-Palmarini, M. (ed), *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Cambridge, MA: Harvard University Press.

Forbus, K. D., Gentner, D., & Law, K. (1994). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19, 141–205.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin*, 105, 387–396.

Gendler, Tamar Szabo. (1998). "Galileo and the Indispensability of Scientific Thought Experiment." *British Journal for the Philosophy of Science* 49: 397–424.

Gendler, Tamar Szabo. (2004). "Thought Experiments Rethought – And Reperceived." *Philosophy of Science* 71: 1152–1163.

Gendler, Tamar Szabo. (2007). Philosophical Thought Experiments, Intuitions, and Cognitive Equilibrium *Midwest Studies in Philosophy, XXXI (2007)*

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy, *Cognitive Science*, 7, 155-70.

Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and cognition* (pp. 195-236). Cambridge, MA: MIT Press.

Gentner, D. (2005). The development of relational category knowledge. In L. Gershkoff-Stowe & D. H. Rakison,(Eds.), *Building object categories in developmental time.* (pp. 245-275). Hillsdale, NJ: Erlbaum.

Gentner, D., and Colhoun, J. (2010). Analogical processes in human thinking and learning. In A. von Müller and E. Pöppel (Series Eds.) and B. Glatzeder, V. Goel, and A. von Müller (Vol. Eds.), *On Thinking: Vol. 2. Towards a Theory of Thinking*. Springer-Verlag Berlin Heidelberg.

Gentner, D., Loewenstein, J., Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology,* 95(2), 393-408.

Gentner, D., Loewenstein, J., Thompson, L., & Forbus, K. (2009) Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. *Cognitive Science*, 3, 1343-1382.

Gentner, D., Rattermann, M. J., and Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology, 25,* 524-575.

Gick, M. L., & K.J. Holyoak. (1980). Analogical problem solving. *Cognitive Psychology*, 12, 306-355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer, *Cognitive Psychology*, 15, 1-38.

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press.

Hogarth, R. M. (2001). *Educating intuition*. University of Chicago Press.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, 15, 332–340.

Horvath, J. (2010). 'How (Not) to React to Experimental Philosophy', *Philosophical Psychology*, 23, 447-480.

Johnson, E. J. (1988). Expertise and decision under uncertainty: Performance and process. In Glaser, Robert, Michelene TH Chi, and M. J. Farr, (eds.) *The nature of expertise*: 209-228.

Kahneman, D. and Klein, G. (2009). ‘Conditions for Intuitive Expertise: A Failure to Disagree’, *American Psychologist*, 64, 6, 515-526.

Klein, G. A. (1998). *Sources of power: How people make decisions*. MIT press.

Kornblith, Hilary, (1998). “The Role of Intuition in Philosophical Inquiry: An Account with No Unnatural Ingredients”, in DePaul and Ramsey (eds.), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, 129–142. Lanham, MD: Rowman and Littlefield.

Lowe, E. J. (2002). *A survey of metaphysics* (Vol. 15). Oxford: Oxford University Press.

Lowenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review, 6*, 586–597.

Machery, E. (2017). *Philosophy within its proper bounds*. Oxford University Press.

Martí, G. (2012). Empirical data and the theory of reference. *Reference and referring: Topics in contemporary philosophy*, 63-82.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Millikan, R. G. (1996). On swampkinds. *Mind & Language*, *11*(1), 103-117.

Neander, K. (1996). Swampman meets swampcow. *Mind & Language*, *11*(1), 118-129.

Nersessian, Nancy. (1993). “In the Theoretician’s Laboratory: Thought Experiments as Mental Modeling”, in *Proceedings from the Philosophy of Science Association*, 291–301.

Nolan, D. (2015). The a posteriori armchair. *Australasian Journal of Philosophy*, *93*(2), 211-231.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 510-520.

Papineau, D. (1996). Doubtful intuitions. *Mind & Language*, *11*(1), 130-132.

Paul, L. A. (2012). Metaphysics as modeling: the handmaiden's tale. *Philosophical Studies*, 160(1), 1-29.

Reiss, J., & Kitcher, P. (2009). Biomedical research, neglected diseases, and well-ordered science. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia*, *24*(3), 263-282.

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

Saariluoma, P. (1990). Apperception and restructuring in chess players' problem solving. In K. J. Gilhooly, M. T. Keana, R. H. Logie & G. Erdos (Eds.), *Lines of thinking: Reflections on the psychology of thought.* (Vol. 2, pp. 41-57). Oxford, England: John Wiley & Sons.

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, naive similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 29, 641–649.

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes*, 53, 252–266.

Sober, E. (2000). Psychological egoism. *The blackwell guide to ethical theory. Blackwell, Malden, MA*, 129-148.

Sober, E., & Wilson, D. S. (1999). *Unto others: The evolution and psychology of unselfish behavior* (No. 218). Harvard University Press.

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

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

Stich, S. (2007). Evolution, altruism and cognitive architecture: a critique of Sober and Wilson's argument for psychological altruism. *Biology & Philosophy*, *22*(2), 267-281.

Swoyer, C. (1999). Explanation and inference in metaphysics. *Midwest Studies in Philosophy 23*, 100–131.

Tetlock, P. E. (2005). *Expert Political Judgment: How Good is It? How can I Know?* Princeton, Princeton University Press.

Thagard, P. (1989). Explanatory coherence. *Behavioral and brain sciences*, *12*(3), 435-467.

Weinberg, Jonathan M. (2007). On how to challenge intuitions empirically without risking scepticism. *Midwest Studies in Philosophy, XXXI*, 318–343.

Weinberg, Jonathan M., Chad Gonnerman, Cameron Buckner, Joshua Alexander. (2010). Are philosophers expert intuiters? *Philosophical Psychology* 23: 331-55.

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

Williamson, Timothy. (2011). Philosophical expertise and the burden of proof. *Metaphilosophy* 42: 215-29.