PhiMiSci
Philosophy and the Mind Sciences

The replication crisis and philosophy

Wesley Buckwalter[a] (wesleybuckwalter@gmail.com)

Abstract
The replication crisis is perceived by many as one of the most significant threats to the reliability of research. Though reporting of the crisis has emphasized social science, all signs indicate that it extends to many other fields. This paper investigates the possibility that the crisis and related challenges to conducting research also extend to philosophy. According to one possibility, philosophy inherits a crisis similar to the one in science because philosophers rely on unreplicated or unreplicable findings from science when conducting philosophical research. According to another possibility, the crisis likely extends to philosophy because philosophers engage in similar research practices and face similar structural issues when conducting research that have been implicated by the crisis in science. Proposals for improving philosophical research are offered in light of these possibilities.

Keywords
Cognitive science · Replication crisis · Methods · Thought experiments · Incentives · Collaboration

1 Introduction

A substantial proportion of published scientific research fails to replicate and is likely unreliable (Ebersole et al., 2016; Klein et al., 2018; Nosek & Lakens, 2014; OSF, 2015; Stroebe, 2019). Several high profile replication attempts have shown that, of the results selected for replication attempts by researchers on various occasions, anywhere from 64% of scientific findings in psychology (OSF, 2015), 39% in experimental economics (Camerer et al., 2016), 89% in preclinical cancer research (Begley & Ellis, 2012; Nosek & Errington, 2017) and 100% of studies involving structural brain-behavior correlations in neuroscience (Boekel et al., 2015) could not be confirmed. Though it is currently unclear to what degree these percentages actually represent replication rates in these fields, concerns over the reliability of and confidence in scientific results are widely shared among scientists, the media, and the general public (Aschwanden, 2019; Baker, 2016). This has led researchers to conclude that "whether 'crisis' is the appropriate term to describe the current

[a]Department of Philosophy, Institute for Philosophy and Public Policy, George Mason University

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193
 https://philosophymindscience.org ISSN: 2699-0369

state or trajectory of science," there is nonetheless, "substantial room for improvement with regard to research practices to maximize the efficiency of the research community's use of the public's financial investment in research" (Munafò et al., 2017, p. 1).

With every crisis, as they say, also comes opportunity. One positive aspect of the replication crisis is the opportunity to study, better understand, and improve upon research practices moving forward. These efforts have largely come from within the fields of psychology and cognitive science. Several psychologists have seized this opportunity by attempting to isolate the causes of the replication crisis and advocate changes to increase reliability and efficiency of research in their field (Ioannidis et al., 2014; Munafò et al., 2017; Romero, 2018; Simmons et al., 2011). These efforts have inspired both methodological and social reforms in psychology, and according to one estimate of average replicability, may have begun to show measurable improvements to published research in social psychological science (Schimmack, 2017).

The opportunity to improve replicable science also extends to researchers working in areas typically thought to be outside of science. Many of the questions raised by the replication crisis are conceptual or philosophical in nature (Fidler & Wilcox, 2018; Romero, 2019). For example, what does it mean to "replicate" a finding, successfully, convincingly, conceptually, or otherwise (Brandt et al., 2014; Hüffmeier et al., 2016; Machery, 2020)? What role do values play in promoting reproducible research, and particularly, values such as openness, trust, civility, and shame in science (Fiske, 2016; Levin & Leonelli, 2017; Wilholt, 2012)? And given all of this, what does it mean to make "progress" in science (Vazire, 2018)? Such questions overlap significantly with foundational research topics in philosophy and philosophers of science may be well positioned to contribute to them.

The overlap between philosophy and science also suggests that philosophers can learn from what is happening in science. Though news of the replication crisis has been dominated by social psychology, all signs indicate that it likely extends to several other fields. According to one hypothesis, while the same problems face many other disciplines, various aspects of the research culture in psychology made the problem easier to detect in that field (Gelman, 2016). For example, it could be that statistical sophistication and transparency made it more likely that the problem would be recognized in psychology. This raises a troubling question: just how far does the replication crisis extend, and could it extend to very different fields across or even outside the sciences?

The aim of this paper is to investigate the possibility that the replication crisis and related challenges to science extend to the field of philosophy. This investigation is motivated by the assumption that just as philosophers can contribute to improving replicable science from a discipline traditionally thought outside of science, so too can scientists contribute to our understanding of philosophical research methods from outside of philosophy. Given the typical objects of philosophical inquiry, this opportunity should be especially welcome. Philosophers

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

often study objects that are intangible, unobservable, or challenging to measure in various respects. This makes it more difficult to directly assess the reliability of research findings in philosophy than it can sometimes be in science. But the crisis in science may serve as a useful model to assess these matters indirectly. If there is significant overlap between fields in some important ways, then it may indicate that the crisis extends to philosophical research. If the crisis does extend to philosophical research, then this might also suggest that similar reforms advocated in science could improve the reliability of research in philosophy, too. For this reason, philosophers might look to the replication crisis not only as an object of research, but also for insights to better understand and improve upon their own activity in light of similarities between what they are doing and what is happening in science.

Here is how the paper will proceed. Section 2 presents several potential indicators or warning signs of an impending crisis in philosophy that motivate improvements to philosophical methods. The next two sections investigate concrete ways that the replication crisis might extend to philosophical research. Section 3 assesses the possibility that a crisis extends to philosophy because philosophers rely on unreplicated or unreplicable evidence when conducting philosophical research. Section 4 assesses the possibility that the crisis extends to philosophy because philosophers engage in similar practices and face similar structural issues as those implicated by the crisis in science. Section 5 proposes several changes to philosophical research practices modeled after some reforms that have improved the reliability of scientific research.

2 Indications of crisis

Before hypothesizing about ways that the replication crisis might extend to philosophy, a more fundamental question arises: are there antecedent reasons to suspect that philosophy is in crisis? After all, it doesn't make much sense to speak of a crisis when there is no cause for alarm and there do not seem to be many philosophical emergencies. Then again, the current problems facing science make a lot more sense in hindsight. Several contributing factors to the replication crisis in science that were widely accepted by researchers only a few short years ago are regarded as completely unacceptable today. But at the time, many scientists openly engaged in and encouraged them. In some literatures, such as ego depletion, for example, foundational theories are now nearly debunked despite widespread support and hundreds of positive results reported in their favor (Hagger et al., 2016). This has led researchers to wonder, "if a large sample pre-registered study found absolutely nothing, how has the ego depletion effect been replicated and extended hundreds and hundreds of times?" and "[…] more sobering still: What other phenomena, which we now consider obviously real and true, will be revealed to be just as fragile?" (Inzlicht, 2016). It is possible that we could also be asking the same questions of philosophy in the not so distant future.

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

Several potential warning signs support this possibility. The following discussion raises several warning signs that collectively suggest that there is room for improvement in philosophical methods. The first warning sign stems from the fact that philosophical research is mired in persistent disagreement. Though there is no established measure of disagreement across the field, it is difficult to deny that disagreement is an inescapable aspect of philosophical activity. If there is research that is advanced by a philosopher, then the chances are good that there is a philosopher who is equally well trained and informed that advances the opposite. Quite often, many of these disagreements are intractable. They are rarely resolved through reflection, discussion, or additional information.

There have been many explanations offered to explain persistent disagreement in philosophy (Beebee, 2018; Daly, 2017; MacBride, 2014). However, one reason disagreement continues unresolved, it is sometimes suggested, is because philosophical researchers rely on different evidence (Beebee, 2018). To a large extent, this evidence consists of intuitions from thought experiments or the judgments that individual philosophers make about real or imagined cases that they construct to support their theories (more below). In metaphysics, for example, researchers have observed several clashes among case judgements in foundational thought experiments:

> You might think that Designed Ernie in Alfred Mele's 'zygote argument' is not morally responsible, and conclude that nonhistoricist compatibilism is false (Mele, 2006, p. 189). Or you might not (Fischer, 2011, p. 271). You might come up with what you take to be a castiron case of the failure of the transitivity of causation (McDermott, 1995, p. 524). Someone else will inevitably disagree (Lewis, 2000, p. 194). You might think you have described two different possible worlds that agree with respect to the distribution of matters of particular fact but disagree with respect to the laws, thus refuting the claim that the latter supervene on the former (Carroll, 1884, pp. 57–68). Or you might not (Beebee, 2000, pp. 586–592). And so on.
> (Beebee, 2018, p. 5)

When this happens, as in metaphysics above, for example, researchers have further observed that philosophers are left with very little recourse over how to proceed:

> We have no such method for resolving metaphysicians' clashes of intuitions. When I say that Designed Ernie is morally responsible for his crime or that in our transitivity case x really was a cause of z, and you disagree, we have no agreed way of reconciling our differences. There are no empirical facts about which you might correct me, in the light of which I would recognise my mistake and change my mind. I have not illegitimately held fixed some facts that I should not have held fixed, or vice versa – not by my lights, anyway. We might try to convince each other otherwise, of course, and sometimes one of us succeeds; but

> often we both fail. Often we both fail simply because each of us is holding fixed some element of our own background philosophical theory that the other rejects. But then our thought experiment serves only to provide an example of the different consequences of our respective theories; it cannot adjudicate between them. (Beebee, 2018, pp. 5–6)

One hypothesis for why intractable disagreement persists is because one philosopher is claiming to offer basic evidence that another philosopher cannot replicate, reproduce, or confirm. In this case, the relevant evidence is a judgment or intuitive reaction from a thought experiment when using the method of cases. If, try as they might, a philosopher cannot reproduce this evidence, while another insists that the evidence exists and supports their theory, then the pair are at an impasse.

Few would claim that researchers must always agree with one another in order for progress to occur. But failure to converge on even basic evidence in core thought experiments when practicing the case method is a sign that philosophical methods are in crisis, especially when there is no way to adjudicate judgments between researchers. Such an impasse leads to several questions reminiscent of those raised in the early days of the scientific crisis. If your peers conduct research that generates results diametrically opposed to your own, and you have every reason to believe that these peers are just as good at conducting research as you are, then it is reasonable to begin questioning the research process. If disagreement persists because two researchers are drawing on different evidence when evaluating thought experiments, then this reasonably calls for a careful investigation into the nature of that evidence. If upon investigation, there is no way to adjudicate the conflicting evidence generated between researchers, then this reasonably calls into question the source of that evidence and points to weaknesses in the research method used to generate it.

A second warning sign stems from the questions of whether philosophy produces a distinctive body of knowledge or makes appreciable progress. Many philosophers are skeptical that it does. According to Thomas Reid, philosophy only manages to "cast a 'darkness visible' on the human faculties, and to disturb the peace and security enjoyed by happier people" (Reid, 1764/1997). William Lycan writes that "philosophical consensus is far more the result of Zeitgeist, fad, fashion, and careerism than of accumulation of probative argument" (Lycan, 2013, pp. 116–117). Still other researchers have claimed either that it is "not clear that the philosophical enterprise has served as a source of knowledge" (Kornblith, 2013, p. 260), or worse, that "there is no information and there are no facts to be learned besides information and facts about what certain people think" in philosophy (Van Inwagen, 2015, p. 11).

When evaluating progress in the field, philosophers often compare the state of philosophical research to that of the natural sciences. For example, it has been suggested that "since science took its modern form in the seventeenth century, it has been one long success story" and that "philosophy compares badly with science on this score" (Papineau, 2017). It is also often suggested that science makes

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

progress because science is self-correcting. The basic idea is that while scientists might make just as many mistakes as other researchers do in the short term, science somehow corrects for this in such a way that promotes truth or instrumental advances in the long term. However, the replication crisis has recently challenged whether science makes significantly more progress than other fields. For example, it is reasonable to question the degree to which science self-corrects for certain types of errors (Romero, 2019). And without successful replication in science, it has even been suggested that, "perpetuated and unchallenged fallacies may comprise the majority of the circulating evidence" (Ioannidis, 2012, p. 645). This sounds a lot like what some skeptics have said about philosophy.

These considerations offer a tentative though suggestive warning that philosophy faces a crisis similar in kind to that of the natural sciences and calls for the improvement of philosophical methods. It could be not that it was wrong to question progress in philosophy, but that it was wrong not to be equally suspicious of scientific progress. The lack of error correction mechanisms needed for the satisfactory promotion of true theories or accumulation of knowledge might well be shared by both fields. In science, the lack of error correction has been tied to concerns over replication and reliability of evidence. Thus, it is reasonable to investigate whether the same features might account for errors that also limit progress or knowledge creation in philosophy.

A third warning sign stems from the growing body of research in cognitive science suggesting that many popular research programs in theoretical philosophy have turned out to be false starts, in some cases because case judgments are not shared or are misunderstood (Buckwalter, 2014; Buckwalter & Turri, 2019; Colaco et al., 2014; Kneer & Machery, 2019; Machery, 2017; Rose et al., 2014, 2017, 2019; Turri et al., 2015; Turri, 2017). This research demonstrates how theorists can be led astray by confounded thought experiments or idiosyncratic understandings of concepts and terminology. In many instances, judgments about cases are either not shared by other researchers or the general public, or have replicated in these ways but are overstated, misdiagnosed, or explained by extraneous factors that were not of philosophical interest.

Two recent developments exemplify these concerns. In epistemology, for example, philosophers have constructed famous pairs of cases manipulating certain variables such as stakes and have claimed that different judgments about them motivate the theory that the word "knows" is a contextually sensitive expression (DeRose, 1992, 2009). In fact, it is often claimed that the effect is so strong that it can reverse the truth of knowledge sentences. When researchers subjected these thought experiments to controlled testing, however, they found little evidence for the predicted effect anywhere in the world (Rose et al., 2019). In a large-scale cross-cultural replication attempt involving forty-five hundred participants in over a dozen countries, researchers failed to replicate the effect at sixteen out of nineteen international research sites. What effects researchers were able to detect were small and amounted to a three percent difference in judgments across conditions.

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

Regardless of whether the effect in question is ultimately detectable in some form or other (Turri et al., 2016; Turri, 2017), this state of affairs suggests that there is room for improvement in investigating it.

In other subfields, such as normative ethics, researchers have partially replicated foundational intuitions and judgments about cases, but in ways that question the development of the resulting philosophical research programs. In one series of experiments, for example, researchers could not detect many of the everyday judgments said to underlie the philosophical puzzle of resultant moral luck (Kneer & Machery, 2019). According to philosophers, this puzzle arises because we are sometimes inclined to judge others for the unlucky consequences of actions even though the results are beyond their ability to fully control (Williams, 1981). When subjecting multiple moral luck cases to controlled experimental testing, however, researchers discovered little evidence for the claim that wrongness, blame, or permissibility judgments differed considerably as a result of lucky or unlucky outcomes. What evidence researchers did find for this effect appeared to be artifacts of the testing situation or perhaps attributable to cognitive biases, such as hindsight bias. At the same time, researchers did find evidence for a substantial effect on other judgments in moral luck cases, such as punishment judgments. If these findings are correct, then they suggest that either moral luck is not a genuine ethical puzzle that arises for many people upon reflection or that while there may be a glimmer of a puzzle here, it has been misunderstood or mischaracterized for several decades.

Whether or not these observations are indicative of a crisis, they indicate room for improvement and warrant further exploration of the possibility. There is persistent philosophical disagreement between researchers that may be perpetuated by differences in case judgements. Longstanding debates about the nature of philosophical progress may point to the need for new methods or procedures to limit the accumulation of errors, as the replication crisis suggests is needed in science. Lastly, replication failure in philosophy may be quite literal. To the extent that case judgments motivating leading theories are not widely shared or unreliably generated, philosophy may share key features of unreplicable science.

3 Philosophy in the age of replication crisis science

Given that there is initial theoretical and practical motivation to investigate the replication crisis in philosophy, what forms could it take? One of the most straightforward ways that the replication crisis extends to philosophy lies in the fact that much research in philosophy relies on unreplicated or unreplicable science. Contrary to how it may sometimes appear, a non-trivial proportion of philosophical research draws on scientific evidence. Philosophers regularly rely on new, interesting, or surprising findings across the sciences to motivate, challenge, or support

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

philosophical theorizing. They also often rely on classic or foundational research in science to do these things, which until recently at least, was considered settled science. However, both kinds of findings are frequently subject to failed replication. Thus, to the extent that philosophers draw on scientific findings in their theorizing, and those findings are implicated by the replication crisis, the replication crisis straightforwardly extends to philosophy.

To what extent does philosophical research rely on evidence from science? Because philosophy is a broad discipline, the answer to this question is bound to vary widely based on the area and research question. Some areas may rely on scientific evidence more than others do or prioritize analytic aspects of various phenomena. Other areas of philosophy, however, rely quite heavily on scientific findings and methods. Some of these areas are necessarily interdisciplinary. Philosophy of science, philosophy of cognitive science, moral psychology, applied social and political philosophy, feminist philosophy, and applied ethics all heavily appeal to developments in various social and natural sciences. This is to be expected of many areas of philosophy that are more likely to address practical philosophical questions that arise in social or political life.

Such appeals are not only restricted to areas of philosophy that study applied questions. Areas of philosophy traditionally associated with or dominated by armchair methods have also dramatically shifted to embrace empirical methods. One particularly striking example of the shift from a priori to empirical methods in a relatively short amount of time has been documented in philosophy of mind (Knobe, 2015). To document this shift, researchers analysed the 397 highest cited articles in the same set of philosophy journals that were published either in the mid-to-late twentieth century (1960–1999), or those published in the early twenty-first century (2009–2013). What researchers found was that there was a change in predominate methods. Papers in the twentieth century were dominated by a priori armchair methods (62.4%). Conversely however, the majority of twentieth century papers relied on empirical research generated in science (61.8%), or contributed new experimental research to the research record (26.8%). Very few papers in the latter period relied on purely a priori methods (11.5%). Researchers also documented a surprising shift in topics of philosophical research during these periods. The earlier period heavily focused on research pertaining to the metaphysics of mind, such the mind-body problem or the nature of content. However, the later period focused on interdisciplinary topics in cognitive science, such as perception, theory of mind, or cognition. These findings suggest a transformation has occurred in analytic philosophy of mind, prioritizing interdisciplinary topics and an engagement with evidence from science.

Appeals to scientific evidence may also extends to several other core areas of philosophy. Often, it is not so much debated whether philosophy relies heavily on science, but whether this development is good or bad. In moral psychology, for example, researchers have argued that popular forms of argument in the field such as debunking arguments "tend to rely on a problematic scientism, privileg-

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

ing scientific causal explanation of targeted ethical or meta-ethical beliefs while ignoring or downplaying important philosophical alternatives" (FitzPatrick, 2018, p. 234). In epistemology, researchers have claimed that "science not only helps us to address the philosophical questions that we had before we became acquainted with scientific advances; it helps us to revise the philosophical questions we ask in light of the better understanding of various phenomena that science provides" (Kornblith, 2018, p. 146). In other areas, researchers lament that philosophers do not rely on science enough. In metaphysics for instance, researchers have argued that analytic metaphysics should be discontinued because "no alternative kind of metaphysics can be regarded as a legitimate part of our collective attempt to model the structure of objective reality" than one that is "radically naturalistic" (Ladyman et al., 2014, p. 1). These observations suggest that scientific evidence is relevant to where philosophy has been, or where it may soon be headed.

Given that philosophical research relies on scientific evidence generated during a replication crisis, to what extent are the findings that philosophers rely on implicated by that crisis? At present writing we do not know the answer to this question. We do not know this because we do not know the true replication rate in science. Successful replication is both bound to fluctuate between scientific fields and vary by individual research questions. Relatively few scientific findings have been subject to registered replication attempts. To further complicate matters, even if findings do replicate, we do not know their true effect size, whether they have been correctly interpreted, or whether they extend from the lab to meaningful contexts in everyday life.

Nonetheless, several high-profile scientific findings of deep philosophical interest have played a significant role in both philosophy and the replication crisis. The Stanford prison experiment, a study frequently discussed in ethics and regarded as a powerful source of evidence for situationism has been called "a lie" (Blum, 2018) involving a "biased and incomplete collection of data" (Le Texier, 2019). The effect of disgust induction on moral judgment, a finding heavily invoked in the metaethics and philosophy of emotion, has been overstated, has failed to replicate, and may be accounted for by publication bias (Ghelfi, 2020; Johnson et al., 2016; Landy & Goodwin, 2015). Relevant to the egoism and altruism debate, the claim that deliberate perspective taking of needy others increases empathetic concern has been heavily challenged (McAuliffe et al., 2020), as has bystander apathy, relevant to the bystander effect (Philpot et al., 2020). It is unlikely that stereotype threat and intelligence mindsets, frequent topics in feminist philosophy and social political philosophy, can explain performance outcomes outside of lab contexts after controlling for publication bias (Bahník & Vranka, 2017; Finnigan & Corker, 2016; Flore & Wicherts, 2015; Shewach et al., 2019). Frequent objects of theorizing in philosophy of mind and philosophy of cognitive science, such as ego depletion (Hagger et al., 2016) or backfire effect (Wood & Porter, 2019) fail to replicate. An extensive meta-analysis appears to undermine previous claims that manipulating free will beliefs is associated with anti-social behavior, such as cheating (Genschow

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

et al., 2022). The implicit association test, a keystone of philosophical theorizing across ethics, metaphysics, and philosophy of mind faces a series of challenges regarding construct validity (Schimmack, 2019) or usefulness (Buckwalter, 2019; Forscher et al., 2019; Machery, 2022). And the finding that intuitions about intentional action are influenced by implicit bias fails to replicate (Klein et al., 2018). These challenges deeply question the starting assumptions of large bodies of literature in philosophy and require significant attention to address.

A final question that remains is to consider how the effects of the scientific crisis might translate to fields typically thought to be outside of science, such as philosophy, when drawing on scientific evidence. One reasonable hypothesis is that the effects will largely be the same between fields. Researchers across these fields ultimately rely on the same body of unreliable evidence to support their theories and will suffer equally when they turn out to be based on unreliable evidence. Another hypothesis, though, is that the effects in philosophy will be worse. There are some important field-specific differences that might magnify the impact of the replication crisis in philosophy. For instance, philosophical research often tends to prioritize extended theorizing. This extended theorizing abstracts away from the specific findings reported to explore the possible philosophical implications or applications of findings. If this is the case, then unreplicable science might license more discussion and speculation in philosophy than it does in science. Extended speculation could magnify the problem by increasing the reach of a single unreplicable result considerably further than data driven science typically allows and by giving a veneer of scientific authority to claims that have not been tested or shown.

Another difference is that many scientific results are not settled and that philosophers have less expertise in evaluating this than professional scientists do. It can be incredibly difficult to accurately assess what a paper has shown, even for those with extensive hands on experience conducting experiments in that research area. But scientists often have better access to experiences, types of background information, and networks between scientists than others do. These things all provide information that often doesn't make it into published papers and that can help contextualize results. Without this access and experience, even the most conscientious researchers are more likely to misinterpret contributions or overlook important red flags. For these reasons, a little bit of unreliable scientific evidence might translate into large effects in philosophy. Large bodies of literature meticulously investigating the possible philosophical implications of unreplicable scientific findings could well be without foundation.

4 Replication and the method of cases

Is the replication crisis only a concern for empirical approaches to conducting philosophical research? It is unsurprising that the replication crisis in science would extend to fields outside of science that sometimes rely on the same evidence from science. Thus, it might be thought that limiting the role or influence

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

of evidence from unreplicable science would prevent the challenge from spreading beyond science. But what of conceptual, analytic, or broadly speaking non-scientific methods in philosophy, could the replication crisis extend to philosophical research produced in these ways regardless of formal contact with scientific evidence? Given the natures of the typical objects of philosophical inquiry, it is difficult to assess this possibility directly. However, it is possible to assess it indirectly, by examining whether the research practices implicated in the scientific crisis are also present in more analytic philosophical methods.

This approach is motivated in part by Bertrand Russell, who argued that philosophy should draw its inspiration from science. By this, however, Russell does not necessarily mean that philosophy should simply inherit the results of science as discussed above, but rather, that it should emulate certain features about the way he perceived science to be optimally conducted:

> Much philosophy inspired by science has gone astray through preoccupation with the results momentarily supposed to have been achieved. It is not results, but methods that can be transferred with profit from the sphere of the special sciences to the sphere of philosophy. What I wish to bring to your notice is the possibility and importance of applying to philosophical problems certain broad principles of method which have been found successful in the study of scientific questions. (Russell, 2008 [1917], 98-99)

The sort of methods that Russell has in mind all generally pertain to a kind of scientific spirit or temperament that prioritizes the use of logic, a dispassionate search for truth, patience, and modesty in the research process. Applying these principles to philosophical research, Russell writes, "is to ensure a progress in method whose importance it would be almost impossible to exaggerate" (ibid., 113-114). Applying Russell's insight in the present case, the thought is that we might be able to continue to improve philosophical research by emulating recent improvements in scientific method. Or put another way, we might be able to isolate weaknesses in conducting philosophical research and improve upon them by observing common points of overlap with methods in science associated with the replication crisis.

Though philosophers might utilize many methods in the course of conducting philosophical research, one natural point of comparison between philosophical and scientific research practices involves the use of the method of cases. The method of cases is a distinctive method in philosophy (Machery, 2017; Strevens, 2019). When applying this method, philosophers plan, construct, and evaluate cases to assess philosophical claims about philosophical phenomena, concepts, common sense, or natural language. Many features of such cases are objects of lengthy philosophical research. In general, however, since the use of cases shares several common features with controlled experiments, it is no surprise that cases are often referred to as "thought experiments".

There are many basic similarities between thought experiments and controlled experiments. When conducting a thought experiment, a philosopher designs a

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

vignette involving a real or imaginative situation. The vignette is designed to test a specific hypothesis, for example, that knowledge is compatible with luck, that justice is repaying debts, that scientific progress can occur without justification, and so on. To test hypotheses about these research questions, philosophers often manipulate factors, for example, by varying whether a protagonist in a vignette acquires information by luck. Philosophers then collect evidence in the form of judgments about the case. If the philosopher is inclined to judge of this situation that, for example, the protagonist has knowledge despite the lucky circumstance, then this is taken as data supporting the original hypothesis that knowledge is compatible with luck. If not, then this is taken as evidence that the hypothesis should be rejected or refined.

This is similar in form to what some scientific researchers do when they conduct controlled experiments, particularly in social psychology. Often, psychologists develop narrative cover stories to use in experimental materials that closely resemble the thought experiments that philosophers use in their papers. The cover stories typically describe ordinary situations that isolate and manipulate variables of psychological interest. These variables frequently overlap with many of the same areas that philosophers study, such as belief, knowledge, morality, intention, or punishment. Sometimes, experimentalists even adopt the exact text of thought experiments originally introduced by philosophers, such as moral dilemmas like trolly cases. Of course, psychologists typically recruit more participants to evaluate cover stories than philosophers do. Psychologists also typically subject those judgments to some kind of statistical analysis beyond armchair reflection. Interestingly, some ways that psychologists have done these things has probably exacerbated the crisis. But in any event, and in this corner of social psychology, at least, the differences to thought experiments in philosophy seems to be a matter of detail rather than kind. And while differences can be substantial when comparing particular research programs in philosophy and psychology, the overlap in the case method motivates the search for additional similarities.

The following is an investigation of many more similarities between research in philosophy using the method of cases and research in social psychology utilizing case-based experiments. The investigation reveals that many of the similarities between these things are problematic. In its use and practice in philosophy, the method of cases embodies several key factors implicated in the replication crisis in science. Reviewed in what follows are some general similarities between thought experiments in philosophy and controlled experiments in science at the forefront of methodological discussions in science.

Sample Size. One key factor perpetuating the replication crisis is low sample size. It is well known that studies with small sample sizes often have low statistical power, which increases the likelihood that measurable effects are not true effects (Button et al., 2013). It is also well known that studies with smaller sample sizes tend to show larger effect sizes and display greater heterogeneity between studies on the same research question (IntHout et al., 2015). Some of the ways that sample

size has been shown to relate to low replication rates in science are mathematical, while others are observational and may vary across research fields. In both ways, small studies are more susceptible to replication failure, false positives, and inflated effect sizes than larger studies. They also tend to exacerbate existing biases that are present when researchers conduct their research.

Philosophers rely on extremely small samples. The method of cases is usually conducted by individual researchers or a small group of collaborators. For these results to be published, the judgments from cases that researchers report must only be shared (or at least not rejected) by a small group, typically consisting of a handful of editors and peer reviewers. With judgments that result from samples this small, however, we simply do not know if they are shared among researchers. If other researchers do share the relevant judgments, we do not know to what strength or degree. To the degree case judgments are reported and validated in this way, the greater the likelihood that the results are false positives, overstated, or do not reflect true discoveries about philosophical phenomena. The clear conclusion to draw from this is that both fields should assess the samples they are drawing from in conjunction with the inferences they are making about the effects they claim to demonstrate.

Reporting and publication bias. A second factor perpetuating the replication crisis involves biased reporting and the publication of research findings. Publication bias, often referred to as the "file drawer problem" occurs when the result of the experiment influences the likelihood that the study will be reported or published (Rosenthal, 1979; Scargle, 2000). According to this problem, we cannot tell within any given research area how many studies have been conducted but not reported. The reason we cannot tell this is because there is a strong preference to only report and publish positive findings. Because we have differential access to positive and negative findings, researchers cannot fully assess the strength of the existing evidence for a research claim. This is troubling and contributes to the replication crisis because a shockingly small number of unpublished results can greatly increase the risk that positive findings are spurious (Rosenthal, 1979).

All indications suggest that philosophical research is extremely susceptible to publication bias. These biases are likely on both the individual and institutional levels. When philosophers use the method of cases in their research, they almost always only report positive evidence. In philosophy, positive evidence consists of judgments that support a desired theoretical application such as motivating a theory or constituting a counterexample. This means that we do not know how many variations of those cases the philosophers tried before getting the desired judgment. Of course, philosophy is different than science in that it sometimes only takes one successful case to prove the point, whereas science is typically interested in establishing central tendencies. Even in these circumstances, however, reporting only successes and burying failures obscures the full picture. When this happens, for example, the research community cannot assess what seemingly innocuous changes were made to cases get the desired result and evaluate the philosophical

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

significance of this. For instance, does the intuition only arise in cases that involve extreme affect, alternate universes, or outer space? It could be valuable to consider this. Neither can we assess how many other philosophers working in the area also tried and failed to construct a case to reach that judgment. If only positive evidence is reported, then we cannot fully evaluate how strong the evidence is for claims made across research areas and potentially valuable information is lost. When it comes to disagreements about famous thought experiments such as Twin Earth cases, for example, philosophers have observed that publications have become "intramural sports among believers" and that "those who do not share the intuition are simply not invited to the games" (Cummins, 1998, p. 116).

It is also less common, by comparison, that cases and case judgments not supporting desired theoretical outcomes are published. Like science, philosophical publishing also prioritizes positive evidence. When using the method of cases in philosophy, this amounts to reaching a desired judgment that either motivates philosophical theories or serves as a counterexample to them. For this reason, there are strong institutional biases against publishing negative findings from the method of cases that do not support one's position. To the extent that publishing dissenting case judgments is discouraged, one of two bad things is likely to happen. Researchers can self-select out of research areas where they are not able to replicate the same evidence everyone else is drawing on when generating theories, which further contributes to publication bias. Or alternatively, researchers can use the method of cases to generate new judgments about unrelated cases that are more conducive to their own preferred theories about the same research question, in which case the cycle is bound to repeat itself.

Lack of replication. A third factor perpetuating the replication crisis is that relatively few scientific findings are subject to replication labor. Replicable science is a cornerstone of reliable science. One way to ensure that science is replicable and reliable is to subject results to systematic replication attempts. Such replication efforts help to rule out systematic error in the research record and improve our confidence in scientific publications. The lack of systematic efforts to do this is arguably the largest factor in explaining why the replication crisis continued as long as it did and was one of the first steps in assessing the damage.

Like the recent picture in science, there have not been enough coordinated replication efforts of case judgments in philosophy. Some replication attempts have recently been carried out cross culturally and this number is growing (Cova et al., 2019; Machery et al., 2017; Rose et al., 2019). For many foundational cases in philosophy, however, we know far too little about the reliability of published findings. We do not know to what degree other philosophers share the same judgments about them, the extent to which these judgments are made by researchers in other fields more broadly, or beyond. As in science, understanding these things is key to understanding and improving the evidence used in philosophical research.

Some aspects of the case method in philosophy may approximate replication but are insufficient. For example, it might be thought that dialectical exchanges

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

between philosophers in published journal articles can sometimes play this role. Philosophy is often conducted through call and response. While this is true, responses are more likely to involve theoretical implications of case judgments or add new case judgements to the research record rather than dispute judgments in particular thought experiments. The reason for this, as noted above is that clashes of intuitions often result in dialectical stalemates. It might also be thought that replication occurs in the classroom, when exposing students to the case method. While it is true that students are exposed to cases in the course of their philosophical education, we do not fully understand what effect this has and whether case judgements are shared. Results are not shared publicly with the research community, which limits the positive effects of this practice with respect to replication.

Insufficient Training. A fourth factor perpetuating the replication crisis is insufficient scientific training. Many have suggested that problems in psychology stem from the fact that students are not well trained in basic concepts and procedures inherent to the method. In psychology, the lack of training involves basic training in statistics, and particularly, things like understanding p-values, effect sizes, and statistical power (Button et al., 2013; Cohen, 1962; Greenland et al., 2016; Sedlmeier & Gigerenzer, 1989). Lack of concern for statistical power and sample size has long been noted as a "remarkable phenomenon" that has been neglected in many "statistics textbooks used in the graduate training of the investigators" (Cohen, 1962, p. 145).

While metaphilosophy is a growing area of philosophical research in which the method of cases is often discussed, it is striking that researchers receive very little practical training in using the method of cases. Particularly, philosophers are not directly trained in how to construct thought experiments and there are few norms for how to evaluate them. This is especially apparent given that thought experiments are often published despite basic weaknesses in their construction and interpretation. For example, they are often long, complicated, and feature strange situations that might well be responsible for judgments orthogonal to the philosophical factors of interest. Cases are also typically published without adequate controls to help isolate the variables of interest. Conclusions are drawn on the basis of a single case, rather than multiple cases that vary in topic or other incidental details. Judgments are also often based on a single question, without considering closely related variables or alternative ways to phrase that question. In some cases, researchers even go so far as to name protagonists in thought experiments intended to produce positive or negative verdicts "Mr. Havit" and "Mr. Nogot", respectively (Lehrer, 1965). These things increase the risk that cases and judgements made about them are not reliable or more widely generalizable. Research would be improved by more formal training in designing and evaluating thought experiments.

Experimenter effects. A fifth factor perpetuating the replication crisis involves the ability of the experimenter to influence the outcome of an experiment. Researchers must make many choices when designing, conducting, and analyzing experiments. Sometimes these choices are innocuous and arbitrary. Other times,

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

freedom or flexibility can be good, insofar as it promotes creativity and exploration. Such choices can profoundly affect the outcome of experimental research (Landy et al., 2020). When left unchecked or improperly acknowledged, however, researcher degrees of freedom can also introduce errors into the collection or analysis of data that biases the results towards a desired hypothesis (Strickland & Suben, 2012; Wicherts et al., 2016). Such degrees of freedom involve aspects of the experimental design, such as wording or measurement choices, data collection and analysis practices, such as insufficient blinding of participants or experimenters, and the manner that results are reported, such as presenting exploratory analyses as confirmatory. Some of these choices increase the likelihood of false positives and decrease replicability.

Philosophical research is unparalleled in the freedom it affords. As in science, freedom in the selection and investigation of research projects can be beneficial. Similarly, however, unchecked degrees of freedom can also increase the prevalence of questionable research practices and unreliable evidence. For example, researchers are free to utilize cases of any length, complexity, or topic, as well as employ any words, phrases, or characters they see fit in designing thought experiments in the method of cases. Because there are few norms when it comes to conducting the method of cases, researchers are also free to present and discuss judgments about the cases in ways that might influence or alter their assessment. These things increase the likelihood that researchers will construct cases to reach a desired conclusion that may not reflect the way the world is organized and that different researchers might reach different findings when considering the same cases.

One representative example involves case judgments in support of leading theories in epistemology. Researchers have shown that subtle confounds in wording and probing in foundational cases can explain judgments about the word "knows" (Turri 2017) and that understanding the mechanisms that cause them can undermine aspects of their theoretical significance (Buckwalter, 2021). In these studies, researchers examined classic pairs of cases used to motivate epistemic contextualism. For example, consider the following low stakes case:

> Keith and his wife Jane are driving home from work on Friday afternoon. They just received a large check from a client, which Keith plans to deposit in their bank account. It is not important for him to deposit the check before Monday: they definitely do have enough money in their account for all their checks to clear. As they drive past the bank, they see that the lines inside are very long. Keith says, 'I hate waiting in line. I'll just come back tomorrow morning instead.' Jane responds calmly, 'This is really not important, but lots of banks are closed on Saturdays. Do you know that our bank is open tomorrow?' Keith answers, 'It was two Saturdays ago that I went to our bank, and it was open. So, yes, I do know that our bank is open tomorrow.' (Turri, 2017, p. 144)

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

And the following high stakes case:

> Keith and his wife Jane are driving home from work on Friday afternoon. They just received a large check from a client, which Keith plans to deposit in their bank account. It is very important for him to deposit the check before Monday: otherwise they won't have enough money in their account for all their checks to clear. As they drive past the bank, they see that the lines inside are very long. Keith says, 'I hate waiting in line. I'll just come back tomorrow morning instead.' Jane responds anxiously, 'This is really very important, and lots of banks are closed on Saturdays. Do you know that our bank is open tomorrow?' Keith answers, 'It was two Saturdays ago that I went to our bank, and it was open. So, no, I don't know that our bank is open tomorrow.' (ibid.)

It is often claimed that cases with this basic structure motivate epistemic contextualism because they demonstrate the effect of contextual standards on judgments about the truth of knowledge sentences (DeRose, 2011). More specifically, they are often taken to show that when a protagonist says "I do know" in low stakes cases and "I don't know" in high stakes case both statements seem true. If the cases are otherwise identical, then it must be shifting epistemic standards that explain these judgments.

What researchers noticed about these thought experiments, however, was that the cases were not otherwise identical. The cases manipulated more than just contextual standards. They also manipulated whether self-attribution or self-denial of knowledge occurred, i.e., whether or not a protagonist says, "I know" or "I don't know". The presence of this additional variable creates a confound that threatens to undermine the explanation of the case judgment. If it is the way the knowledge statement is phrased, rather than the contextual standard being manipulated between cases, then this challenges the evidence for the effect of contextual standards. When subjected to controlled testing, this is exactly what researchers found (Turri, 2017, Experiment 2). The difference in the way knowledge statements were phrased produced the agreement patterns predicted by contextualism, even when stakes do not vary. In short, judgments were being caused by seemingly innocuous or incidental features introduced into thought experiments by researchers rather than factors of theoretical significance.

Hidden Moderators. A fifth factor perpetuating the replication crisis involves unaccounted for features of an experiment that contribute to its result. More specifically, the term "hidden moderators" is often used to describe unobserved or unmeasured factors that influence effects. Out there in the world, of course, there are a lot of potential moderators that could impact experimental findings as a result of the context in which any one experiment was conducted. This kind of contextual sensitivity might involve, for example, features of the participants, such as culture or native language, the times or settings in which the experiment took place, or design features of the experiment itself, such as details of stimuli or probing method.

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

Insufficient sensitivity to these potential moderators increases the likelihood that a given finding is overstated or is not generalizable to new experimental contexts. It is also possible, in some cases, that these effects explain why findings are either not replicable or might not replicate on some experimental occasions (Van Bavel et al., 2016).

The possibility of hidden moderators may constitute one of the most significant challenges to the method of cases. The reason for this is simple. We do not know nearly enough about the individual features of a thought experiment that are responsible for the judgments that we make about it. Quite often, it is simply taken for granted that we do. Almost always, the research assumption is that whatever judgments thought experiments elicit are caused by what the researcher intended the thought experiment to assess. If a thought experiment is constructed to assess the relationship between knowledge and luck, for example, it is assumed that whatever judgment it elicits is about luck and the direct effect that luck has on knowledge judgments. But this is little more than an assumption. It is only an assumption because the pathways or mechanisms that generate philosophical judgments are often complex, interact with several variables or combinations of variables, and are not fully accessible by introspection. As in science, this raises the possibility that judgments about cases are partially explained by previously unaccounted for factors that alter or constrain their theoretical significance.

Some research in experimental cognitive science and experimental philosophy suggests that this may be common (Buckwalter, 2021; Buckwalter & Turri, 2015; Chituc et al., 2016; Kneer & Machery, 2019; Rose et al., 2017; Turri, 2017). We often misdiagnose the reasons for judgments or misrepresent their causal structure when utilizing the method of cases. One representative example comes from research in metaphysics and action theory concerning intuitive free will judgments. Case judgments have played a significant role in the free will literature, with some philosophers claiming that compatibilism is natural (Fischer & Ravizza, 1998; Nahmias, 2014), while others claiming incompatibilism is natural (Pereboom, 2014; Strawson, 1986). Most of the evidence for natural compatibilism or incompatiblism comes from claims about free will and responsibility judgments made in deterministic thought experiments. Consider, for example, the following case:

> Imagine that in the next century we discover all the laws of nature, and we build a supercomputer which can deduce from these laws of nature and from the current state of everything in the world exactly what will be happening in the world at any future time. It can look at everything about the way the world is and predict everything about how it will be with 100% accuracy. Suppose that such a supercomputer existed, and it looks at the state of the universe at a certain time on March 25th, 2150 A.D., twenty years before Jeremy Hall is born. The computer then deduces from this information and the laws of nature that Jeremy will definitely rob Fidelity Bank at 6:00 PM on January 26th, 2195. As

> always, the supercomputer's prediction is correct; Jeremy robs Fidelity Bank at 6:00 PM on January 26th, 2195. (Nahmias et al., 2005, p. 559)

Given that the thought experiment stipulates that determinism is true, if we intuitively judge that Jeremy acted of his own free will or that Jeremy was morally responsible after considering it, then this demonstrates that determinism is naturally compatible with free will or moral responsibility.

However it is possible to question whether judgments about free will or responsibility support natural compatibilism simply because they follow from reading a case involving determinism. For case judgments to provide evidence that either compatibilism or incompatiblism is natural, it is not enough, it might be thought, to simply make judgments about freedom or responsibility after reading deterministic cases. It is essential that such judgments are also properly responsive to the deterministic nature of the scenarios. If judgments are made without such responsiveness, then it is unclear what evidence they contribute to the debate about whether compatibilism or incompatiblism is natural.

This is precisely what researchers have shown. More specifically, researchers have shown that some judgments in classic free will cases used to motivate compatibilism between determinism, freedom, and moral responsibility are made without tracking the deterministic features of these scenarios (Nadelhoffer et al., 2020; Rose et al., 2017). Instead, researchers have shown that many of the free will and responsibility judgments taken as evidence for natural compatibilism can be explained by commitments to indeterminism about human actions and decision making. Despite what the scenario might say above for instance, people judge that there is a slight chance that Jeremy would not rob the bank as the computer predicted he would. Even though Jeremy actually did what the computer predicted he would do, the thinking seems to go, it was possible for Jeremy to do something else instead. In short, indeterministic commitments were hidden or unmeasured when processing deterministic features of the cases, and the unmeasured effect of these things may explain case judgments. This threatens to undermine their evidential role in theorising about natural compatibilism and raises larger questions about the use of the case method without controlling for such effects.

A related but distinct body of examples also suggests that judgments from the method of case may be moderated by demographics, order effects, framing effects, and situational cues. For example, researchers have shown well-replicated effects of culture or heritable personality traits on philosophical case judgments in philosophy of language and action theory (Beebe & Undercoffer, 2016; Feltz & Cokely, 2012; Machery et al., 2004). Researchers have shown that presentation, framing, and order effects persist in foundational case judgments in ethics, such as trolley problem and Asian disease scenarios, with or without professional philosophical training (Liao et al., 2012; Schwitzgebel & Cushman, 2015; Wiegmann et al., 2012). These things are important to consider and can be difficult to detect when conducting thought experiments.

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

One of the most comprehensive challenges to the method of cases to date is due to Edouard Machery (Machery, 2017). In this book length assessment to the case method, Machery candidly assesses the reliability of effects of culture, gender, age, personality, order, and framing on case judgments. While some effects are more reliable than others, the result of this examination overall is that "nearly all the cases that have been examined are influenced either by demographic or by presentation variables, and this influence is frequently large," (ibid. p. 88). Machery goes on to conclude that, "variation and instability (due to presentation variables such as different framings or different orders of presentation) are thus both substantial and widespread," (ibid). According to a separate analysis of these differences, more than 90 studies to date involving over 75,000 participants in experimental cognitive science have reported demographic variation in judgments to philosophical cases (Stich & Machery, 2022). While there is currently some debate regarding the frequency or strength to which philosophical intuitions vary by demographic groups (Knobe, 2019), the possibility that demographic or presentation effects impact judgments when using the method of cases cannot be ignored. This must be considered and measured responsibly as a matter of course.

Incentive Structure. A seventh factor perpetuating the replication crisis involves the way in which conducting inquiry in science is socially organized. Researchers in science work within a highly institutionalized system of academic credit and reward (May, 2021). Rewards in science such as placement and academic promotion require publishing. To a large extent, publishing requires positive findings. In turn, the pressure to generate positive findings increases the likelihood that the findings that are reported are not reliable or replicable. This situation has led researchers to conclude that "the most powerful incentives in contemporary science actively encourage, reward and propagate poor research methods" and that this process drives the "natural selection of bad science" (Smaldino & McElreath, 2016, p. 2).

Science, of course, is not the only field susceptible to research incentives. Seemingly all academic researchers are incentivised toward maximizing credit and reward in their respective fields. However, the effects of this may be especially pronounced for philosophy and the method of cases. Philosophy also requires publishing for placement and career advancement. And one of the main ways that philosophers build careers in philosophy is by generating examples that question or motivate theories. To generate these things, philosophers often rely on judgments from the method of cases. Thus, they are strongly incentivised to create thought experiments that yield the desired judgments. But unlike controlled experiments, there is a relative lack of procedures, norms, or oversight in the use of thought experiments. Moreover, solitary inquiry and single-authored publications have traditionally been the norm in philosophy. Co-authored publications are sometimes even discouraged or viewed as less valuable contributions in philosophy, which is virtually the opposite of science. One consequence of these practices for the method is that individual researchers often end up being the sole

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

participants of the thought experiments they've created for the purposes of supporting or challenging the theories they wish to support or challenge and the only data generated are their own best judgments about whether they've been successful. This encourages the use of poorly designed thought experiments to produce the desired judgments.

In summary, there are several similarities between the use of controlled experiments in social science and thought experiments in philosophy that overlap significantly with the causes of the replication crisis in science. Both controlled experiments and thought experiments have utilized small samples that increase the likelihood that observed effects are false, overstated, or amplify existing biases in research practices. Lack of replication and biased reporting perpetuate the use of unreliable evidence and limit our understanding of the total evidence. Experimenter effects brought on by rampant researcher degrees of freedom and hidden moderators abound. Lastly, researchers in both fields receive insufficient training in central aspects of the methods and are strongly incentivised to embrace methods that produce desired findings. These similarities between core aspects of the method of cases and methods in social psychological science are suggestive of a similar crisis in philosophy. Indeed, perhaps one way to view the factors that caused the replication crisis in science is that they were improvements to the philosophical method of cases that did not go far enough. These improvements also made visible the problems in the way the method of cases had traditionally been practiced.

5 Proposals for reliable philosophical research

Given that there are some initial warning signs in philosophy, it is reasonable to begin investigating the possibility that the replication crisis and related challenges to the reliability or credibility of science might extend to philosophy. One possibility is that this happens because philosophy relies on evidence generated by replication-crisis-era science. A second possibility is that this happens because some philosophical research methods overlap with structural factors that give rise to unreliable research. Given these possibilities, it is reasonable to explore ways that the costs of the replication crisis might be avoided in philosophy by seizing the opportunity to improve philosophical research. Fortunately, several proposals have been made for improving the replicability of social psychological science (Munafò et al., 2017). Given overlap between fields when it comes to evidence or methods, such proposals might also improve the reliability and efficiency of philosophical research. And while there is a lengthy debate in metaphilosophy about the products of the method of cases and the degree to which they constitute good evidence, the likelihood that they can is increased by adopting the following reforms.

One proposal is to increase sample size through collaboration, group inquiry, and ideally, inquiry led by diverse groups. Diversity and inclusion are widely dis-

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

cussed factors for improving professional philosophy. There is agreement that diversity and inclusion are not only paramount for just and productive inquiry, but also that these things can decrease the effect of individual biases within research communities (Longino, 1990). Prior research has also suggested that collaboration among diverse groups is valued in lay populations when choosing to engage with and study philosophical questions (Buckwalter & Turri, 2016). The present discussion contributes a more specific way that diversity and inclusion are paramount. By including a wider range of researchers into the research process, it increases the chances that case judgments are repeatable and reliable. When the group members have different disciplinary backgrounds, this can also have a training effect and balance out the lack of scientific literacy.

A second proposal is to increase training in both scientific literacy and in using the method of cases. "As the devil can quote Scripture", some researchers have noted, "so the philosopher can quote science" (Russell, 2008 [1917], pp. 43–44). If a large percentage of evidence in philosophy is empirical in nature, then philosophers require the ability to assess it. Some efforts to improve training in some subfields, such as in moral psychology, for instance are underway but have not yet been widely adopted (Machery & Doris, 2017). Doing so lowers the likelihood that philosophers or even entire subfields are duped by misinterpreted or questionable evidence. Improving scientific literacy among philosophers is especially important when training and mentoring students, who might otherwise be encouraged to spend years of their lives on a thesis that is fundamentally motivated by results of no scientific merit. Likewise, additional training is required for conducting the method of cases. This should involve basic training in the construction of cases, use of controls and multiple cover stories, and the unbiased presentation and assessment of cases (Buckwalter, in press).

A third proposal is to encourage reporting and publication of negative findings relevant to philosophical research. Within philosophy, what "negative findings" amount to might take many forms. One idea is to incentivise the publication of papers that grapple with failed arguments for prized claims. One classic example of this comes from the paper "Why is Belief Involuntary" by Jonathan Bennett. The paper begins as follows:

> This paper will present a negative result - an account failure to explain why belief is involuntary. When I announced my question a year or so ahead of time, I had a vague idea it might be answered, but I cannot make it work out. Necessity, this time, has not given birth to invention. (Bennett, 1990, p. 87)

The idea here is that exploring why our arguments for conclusions we like fail can be beneficial to research, especially for claims in philosophy that are currently popular, such as the claim that belief is involuntary.

A different conception of publishing negative results involves changing incentives to encourage more reporting of unsuccessful uses of the case method. If a philosopher has a different intuition or diagnosis of the exact same thought ex-

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

periment, they should be encouraged to publish this finding. But more than just registering disagreement with case judgments from cases already in the published literature, philosophers might also be incentivised to report instances in which they could not build satisfactory cases to yield the case judgments that would generate a problem for a theory that they do not like or support one that they do. For instance, perhaps the researcher can only get the case to work in certain hypothetical situations that are not of central philosophical concern. Sharing this information provides valuable insights into the process of case construction. For example, this might reveal confounding variables that would have otherwise impacted case judgments and distract us from the philosophical research question being investigated.

One way that scientists have helped encourage reporting and transparency is through the use of preregistration and it is interesting to explore how this practice might potentially be used in other fields. Preregistration involves specifying a research plan in a registry in advance of conducting a study (Nosek Brian et al., 2018). Of course, journal articles in philosophy often do not clearly separate methods from results sections and philosophical research does not always proceed by testing hypotheses with empirical studies. Nonetheless, preregistration may still benefit researchers by improving theoretical aspects of research projects (Sarafoglou et al., 2022). For example, one thing preregistration forces researchers to do is to define primary and any secondary research questions and to clearly articulate research hypotheses. This practice can be incredibly valuable when structuring and honing research projects and to execute those research projects more efficiently. Preregistration also forces researchers to specify an analysis plan, which in philosophy, might encourage more careful reflection about methods and evidence. For example, this might involve determining ahead of time whether evidence will come from social observation, personal experience, logical inference, historical analysis, or the method of cases, as well as encourage further reflection about the strengths and weaknesses of these sources for answering the central research question that has been identified. These may be things that early career researchers in particular could benefit from, especially if they involve input from peers at an early stage as to whether a research question, argument, or vignette is well specified.

Lastly, judgments from broader samples of researchers or the community should also be encouraged. Thankfully, philosophy journals have indicated some willingness to publish replication attempts of foundational case judgments and the reinterpretation of past findings (Kim & Yuan, 2015; Sytsma & Livengood, 2011; Turri, 2014). Likewise, we might encourage additional review articles that document the strength of existing evidence in a research area. More broadly, a systematic repository of philosophically relevant replication attempts across the sciences would be extremely valuable in raising awareness about the reliability of scientific findings for philosophical theorizing.

A fourth proposal for improving philosophical research is to limit the influence of experimenter effects and investigate potential moderators for philosoph-

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

ical judgments when practicing the method of cases. To do so requires a better scientific understanding of the psychological processes and mechanisms that underly case judgements. This necessity was partially foreseen by David Hume, who wrote that "the only expedient, from which we can hope for success in our philosophical researches" is to study "human nature itself; which being once masters of, we may every where else hope for an easy victory" (Hume, 1978, pp. I:6–8). In the same passage, Hume tells us that "there is no question of importance, whose decision is not comprised in the science of man; and there is none, which can be decided with any certainty, before we become acquainted with that science" (ibid). Applying Hume's insight to the present discussions, scientific research into the reasons philosophers make the judgments that they do goes a long way to improving our understanding of philosophical activity and research outcomes.

One way to interpret Hume's vision is to fully embrace the methods of science when utilizing the case method in philosophy by conducting controlled experiments. There are some initial indications that doing so objectively improves the likelihood that the method produces reliable philosophical evidence. For example, one recent audit found that nearly three out of four studies in the field of experimental philosophy selected for attempted replication were successful, suggesting a relatively high replication rate compared to other areas of social science (Cova et al., 2021). A separate audit found that statistical irregularities in experimental philosophy may be lower than those in other fields (Colombo et al., 2018). Of course, just because a judgment is shared does not necessarily mean that it is philosophically significant. But these audits do suggest that adopting experimental techniques with the reforms above in mind can be an effective way to improve the reliability of philosophical evidence.

Another way to interpret Hume is to emphasize the important advantages of interdisciplinary research that includes both strong conceptual and experimental foundations. This suggests a fifth proposal, which is to encourage interdisciplinary research with these components. For example, the topic of mind wandering and the phenomenon of streams of consciousness is of both deep philosophical and scientific interest. In recent years, however, it has become clear that an efficient way to make progress understanding this phenomenon involves a combination of conceptual clarity in philosophy in defining central terms, as well as rigorous data collection concerning the neural correlates, ordinary concepts, experiences, and phenomenology of mind wandering (Irving & Glasser, 2020; Mills et al., 2018). To cite another example, foundational questions in philosophy of mind involve the relationships between memory, self-knowledge, and group minds. Recent research suggests that combining conceptual and experimental research methods on autobiographical recall clarifies the ways that collaborative processes impact the quality of certain memories (Selwood et al., 2020). It is reasonable to suppose that grounding research in strong conceptual frameworks as well as experimental reforms motivated by crisis science may increase the reliability of research. Future

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

research might profitably explore the relationship between replication rates and this kind of interdisciplinary inquiry.

In the spirit of interdisciplinarity, a final question remains as to the scope of the argument and the impact of the crisis. Philosophy is not the only field outside of what is typically considered science that either relies on scientific evidence or is subject to the structural pressures of conducting research implicated by the replication crisis. Further study may reveal that the present discussion is but one case study in documenting a problem that generalizes from science to several other disciplines across the arts and humanities. Researchers in these fields may also begin to examine the impact of the scientific crisis on their scholarly activity to improve research.

Acknowledgments
Thanks to Carolyn Buckwalter, David Liggins, Fraser MacBride, John Turri, and Jennifer Windt for comments on previous drafts of this manuscript.

References

Aschwanden, C. (2019). *We're all "p-hacking" now: An insiders' term for scientific malpractice has worked its way into pop culture. Is that a good thing?* www.wired.com/story/were-all-p-hacking-now/

Bahník, Š., & Vranka, M. A. (2017). Growth mindset is not associated with scholastic aptitude in a large sample of university applicants. *Personality and Individual Differences*, *117*, 139–143. https://doi.org/10.1016/j.paid.2017.05.046

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, 452–454. https://doi.org/10.1038/533452a

Beebe, J. R., & Undercoffer, R. (2016). Individual and cross-cultural differences in semantic intuitions: New experimental findings. *Journal of Cognition and Culture*, *16*(3-4), 322–357. https://doi.org/10.1163/15685373-12342182

Beebee, H. (2000). The non-governing conception of laws of nature. *Philosophy and Phenomenological Research*, *61*(3), 571–594. https://doi.org/10.2307/2653613

Beebee, H. (2018). Philosophical scepticism and the aims of philosophy. *Proceedings of the Aristotelian Society*, *118*(1), 1–24. https://doi.org/10.1093/arisoc/aox017

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. https://doi.org/10.1038/483531a

Bennett, J. (1990). Why is belief involuntary? *Analysis*, *50*(2), 87–107. https://doi.org/10.2307/3328852

Blum, B. (2018). *The lifespan of a lie.* https://gen.medium.com/the-lifespan-of-a-lie-d869212b1f62

Boekel, W., Wagenmakers, E. J., Belay, L., Verhagen, J., Brown, S., & Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, *66*, 115–133. https://doi.org/10.1016/j.cortex.2014.11.019

Brandt, M. J., Ijzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & Veer, A. van 't. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Buckwalter, W. (2014). Factive verbs and protagonist projection. *Episteme*, *11*(4), 391–409. https://doi.org/10.1017/epi.2014.22

Buckwalter, W. (2019). Implicit attitudes and the ability argument. *Philosophical Studies*, *176*(11), 2961–2990. https://doi.org/10.1007/s11098-018-1159-7

Buckwalter, W. (2021). Error possibility, contextualism, and bias. *Synthese*, *198*, 2413–2426. https://doi.org/10.1007/s11229-019-02221-w

Buckwalter, W. (in press). A guide to thought experiments in epistemology. In E. Sosa, M. Steup, J. Turri, & B. Roeber (Eds.), *Contemporary debates in epistemology, 3rd edition.* Wiley-Blackwell.

Buckwalter, W., & Turri, J. (2015). Inability and obligation in moral judgment. *PLoS One*, *10*(8), 212–222. https://doi.org/10.1371/journal.pone.0136589

Buckwalter, W., & Turri, J. (2016). Perceived weaknesses of philosophical inquiry: A comparison to psychology. *Philosophia*, *44*(1), 33–52. https://doi.org/10.1007/s11406-015-9680-9

Buckwalter, W., & Turri, J. (2019). Moderate scientism in philosophy. In J. de Ridder, R. Peels, & R. van Woudenberg (Eds.), *Scientism: Prospects and problems* (pp. 280–300). Oxford University Press.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *25;351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Carroll, J. W. (1884). *Laws of nature*. Cambridge University Press.

Chituc, V., Henne, P., Sinnott-Armstrong, W., & De Brigard, F. (2016). Blame, not ability, impacts moral "ought" judgments for impossible actions: Toward an empirical refutation of "ought" implies "can". *Cognition*, *150*, 20–25. https://doi.org/10.1016/j.cognition.2016.01.013

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Colaco, D., Buckwalter, W., Stich, S., & Machery, E. (2014). Epistemic intuitions in fake-barn thought experiments. *Episteme*, *11*(2), 199–212. https://doi.org/10.1017/epi.2014.7

Colombo, M., Duev, G., Nuijten, M. B., & Sprenger, J. (2018). Statistical reporting inconsistencies in experimental philosophy. *PLOS One*, *13*(4), e0194360. https://doi.org/10.1371/journal.pone.0194360

Cova, F., Olivola, C. Y., Machery, E., Stich, S., Rose, D., Alai, M., Angelucci, A., Berniūnas, R., Buchtel, E. E., Chatterjee, A., Cheon, H., Cho, I.-R., Cohnitz, D., Dranseika, V., Lagos, Á. E., Ghadakpour, L., Grinberg, M., Hannikainen, I., Hashimoto, T., … Zhu, J. (2019). De pulchritudine non est disputandum? A cross-cultural investigation of the alleged intersubjective validity of aesthetic judgment. *Mind & Language*, *34*(3), 317–338. https://doi.org/10.1111/mila.12210

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai van Dongen, N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., … Zhou, X. (2021). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, *12*(1), 9–44. https://doi.org/10.1007/s13164-018-0400-9

Cummins, R. (1998). Reflections on reflective equilibrium. In M. DePaul & W. Ramsey (Eds.), *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry* (pp. 113–127). Rowman; Littlefield.

Daly, C. (2017). II—Persistent philosophical disagreement. *Proceedings of the Aristotelian Society*, *117*(1), 23–40. https://doi.org/10.1093/arisoc/aow020

DeRose, K. (1992). Contextualism and knowledge attributions. *Philosophy and Phenomenological Research*, *52*(4), 913–929. https://doi.org/10.2307/2107917

DeRose, K. (2009). *The case for contextualism*. Oxford University Press.

DeRose, K. (2011). Contextualism, contrastivism, and X-Phi surveys. *Philosophical Studies*, *156*(1), 81–110. https://doi.org/10.1007/s11098-011-9799-x

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Feltz, A., & Cokely, E. T. (2012). The philosophical personality argument. *Philosophical Studies*, *161*(2), 227–246. https://doi.org/10.1007/s11098-011-9731-4

Fidler, F., & Wilcox, J. (2018). Reproducibility of scientific results. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility/

Finnigan, K. M., & Corker, K. S. (2016). Do performance avoidance goals moderate the effect of different types of stereotype threat on women's math performance? *Journal of Research in Personality*, *63*, 36–43. https://doi.org/10.1016/j.jrp.2016.05.009

Fischer, J. M. (2011). The zygote argument remixed. *Analysis*, *71*(2), 267–272. https://doi.org/10.1093/analys/anr008

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: An essay on moral responsibility*. Cambridge University Press.

Fiske, T., Susan. (2016). A call to change science's culture of shaming. *APS Observer*, *29*(9).

FitzPatrick, W. (2018). Cognitive science and moral philosophy: Challenging scientistic overreach. In J. de Ridder, R. Peels, & R. van Woudenberg (Eds.), *Scientism: Prospects and problems* (pp. 233–257). Oxford University Press. https://doi.org/10.1093/oso/9780190462758.003.0011

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology*, *53*(1), 25–44. https://doi.org/10.1016/j.jsp.2014.10.002

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, *117*(3), 522–559. https://doi.org/10.1037/pspa0000160

Gelman, A. (2016). *Why does the replication crisis seem worse in psychology?* Slate. https://slate.com/technology/2016/10/why-the-replication-crisis-seems-worse-in-psychology.html

Genschow, O., Cracco, E., Schneider, J., Protzko, J., Wisniewski, D., Brass, M., & Schooler, J. W. (2022). Manipulating belief in free will and its downstream consequences: A meta-analysis. *Personality and Social Psychology Review*, 10888683221087527. https://doi.org/10.1177/10888683221087527

Ghelfi, C. D. C., Eric. (2020). Reexamining the effect of gustatory disgust on moral judgment: A multi-lab direct replication of eskine, kacinik, and prinz (2011). *Advances in Methods and Practices in Psychological Science*, 3–23. https://doi.org/10.1177/2515245919881152

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., Ridder, D. T. D. D., Dewitte, S., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology*, *66*, 81–92. https://doi.org/10.1016/j.jesp.2015.09.009

Hume, D. (1978). *A treatise of human nature*. Oxford University Press.

IntHout, J., Ioannidis, J. P., Borm, G. F., & Goeman, J. J. (2015). Small studies are more heterogeneous than large ones: A meta-meta-analysis. *Journal of Clinical Epidemiology*, *68*(8), 860–869. https://doi.org/10.1016/j.jclinepi.2015.03.017

Inzlicht, M. (2016). *Reckoning with the past*. http://michaelinzlicht.com/getting-better/2016/2/29/reckoning-with-the-past

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*(6), 645–654. https://doi.org/10.1177/1745691612464056

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241. https://doi.org/10.1016/j.tics.2014.02.010

Irving, Z. C., & Glasser, A. (2020). Mind-wandering: A philosophical guide. *Philosophy Compass*, *15*(1), e12644. https://doi.org/10.1111/phc3.12644

Johnson, D. J., Wortman, J., Cheung, F., Hein, M., Lucas, R. E., Donnellan, M. B., Ebersole, C. R., & Narr, R. K. (2016). The effects of disgust on moral judgments: Testing moderators. *Social Psychological and Personality Science*, *7*(7), 640–647. https://doi.org/10.1177/1948550616654211

Kim, M., & Yuan, Y. (2015). No cross-cultural differences in gettier car case intuition: A replication study of weinberg et al. 2001. *Episteme*, *12*(3), 355–361. https://doi.org/10.1017/epi.2015.17

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, *182*, 331–348. https://doi.org/10.1016/j.cognition.2018.09.003

Knobe, J. (2015). Philosophers are doing something different now: Quantitative data. *Cognition*, *135*, 36–38. https://doi.org/10.1016/j.cognition.2014.11.011

Knobe, J. (2019). Philosophical intuitions are surprisingly robust across demographic differences. *Epistemology & Philosophy of Science*, *56*(2), 29–36. https://doi.org/10.5840/eps201956225

Kornblith, H. (2013). Is philosophical knowledge possible? In D. E. Machuca (Ed.), *Disagreement and skepticism* (pp. 260–276). Routledge.

Kornblith, H. (2018). Philosophy, science, and common sense. In J. de Ridder, R. Peels, & R. van Woudenberg (Eds.), *Scientism: Prospects and problems* (pp. 127–148). Oxford University Press. https://doi.org/10.1093/oso/9780190462758.003.0006

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193

Ladyman, J., Collier, J. G., Ross, D., & Spurrett, D. (2014). *Every thing must go: Metaphysics naturalized*. Oxford University Press.

Landy, J. F., & Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science*, *10*(4), 518–536. https://doi.org/10.1177/1745691615583128

Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., Bergh, D. van den, Marsman, M., Derks, K., Wagenmakers, E. J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., … Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*(5), 451–479. https://doi.org/10.1037/bul0000220

Le Texier, T. (2019). Debunking the Stanford prison experiment. *American Psychologist*, *74*(7), 823–839. https://doi.org/10.1037/amp0000401

Lehrer, K. (1965). Knowledge, truth and evidence. *Analysis*, *25*(5), 168–175. https://doi.org/10.2307/3326431

Levin, N., & Leonelli, S. (2017). How does one "open" science? Questions of value in biological research. *Science, Technology, & Human Values*, *42*(2), 280–305. https://doi.org/10.1177/0162243916672071

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, *97*(4), 182–197. https://doi.org/10.2307/2678389

Liao, S. M., Wiegmann, A., Alexander, J., & Vong, G. (2012). Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology*, *25*(5), 661–671. https://doi.org/10.1080/09515089.2011.627536

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.

Lycan, W. G. (2013). On two main themes in Gutting's *What philosophers know*. *Book Symposium, Southern Journal of Philosophy*, *51*(1), 112–120. https://doi.org/10.1111/sjp.12003

MacBride, F. (2014). Analytic philosophy and its synoptic commission: Towards the epistemic end of days. *Royal Institute of Philosophy Supplement*, *74*, 221–236. https://doi.org/10.1017/S1358246114000095

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

Machery, E. (2020). What is a replication? *Philosophy of Science*, *87*(4), 545–567. https://doi.org/10.1086/709701

Machery, E. (2022). Anomalies in implicit attitudes research. *Wiley Interdisciplinary Reviews: Cognitive Science*, *13*(1), e1569. https://doi.org/10.1002/wcs.1569

Machery, E., & Doris, J. M. (2017). An open letter to our students: Doing interdisciplinary moral psychology. In B. G. Voyer & T. Tarantola (Eds.), *Moral psychology: A multidisciplinary guide* (pp. 119–143). Springer International Publishing. https://doi.org/10.1007/978-3-319-61849-4_7

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition*, *92*(3), 1–12. https://doi.org/10.1016/j.cognition.2003.10.003

Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N., & Hashimoto, T. (2017). Gettier across cultures. *Noûs*, *51*(3), 645–664. https://doi.org/10.1111/nous.12110

May, J. (2021). Bias in science: Natural and social. *Synthese*, *199*(1), 3345–3366. https://doi.org/10.1007/s11229-020-02937-0

McAuliffe, W. H. B., Carter, E. C., Berhane, J., Snihur, A. C., & McCullough, M. E. (2020). Is empathy the default response to suffering? A meta-analytic evaluation of perspective taking's effect on empathic concern. *Personality and Social Psychology Review*, *24*(2), 141–162. https://doi.org/10.1177/1088868319887599

McDermott, M. (1995). Redundant causation. *The British Journal for the Philosophy of Science*, *46*(4), 523–544. https://doi.org/10.1093/bjps/46.4.523

Mele, A. (2006). *Free will and luck*. Oxford University Press.

Mills, C., Raffaelli, Q., Irving, Z. C., Stan, D., & Christoff, K. (2018). Is an off-task mind a freely-moving mind? Examining the relationship between different dimensions of thought. *Consciousness and Cognition*, *58*, 20–33. https://doi.org/10.1016/j.concog.2017.10.003

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Nadelhoffer, T., Rose, D., Buckwalter, W., & Nichols, S. (2020). Natural compatibilism, indeterminism, and intrusive metaphysics. *Cognitive Science*, *44*(e12873). https://doi.org/10.1111/cogs.12873

Nahmias, E. (2014). Is free will an illusion? Confronting challenges from the modern mind sciences. In W. Sinnott-Armstrong (Ed.), *Moral psychology, vol. 4: Freedom and responsibility* (pp. 2–56). MIT Press.

Nahmias, E. A., Morris, S. G., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology*, *18*(5), 561–584. https://doi.org/10.1080/09515080500264180

Nosek, B. A., & Errington, T. M. (2017). Making sense of replications. *eLife*, *6*, e23383. https://doi.org/10.7554/eLife.23383

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Nosek Brian, A., Ebersole Charles, R., DeHaven Alexander, C., & Mellor David, T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

OSF. (2015). *Estimating the reproducibility of psychological science*. *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Papineau, D. (2017). Is philosophy simply harder than science? *The Times Literary Supplement*, *December 2, 2019*.

Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford University Press.

Philpot, R., Liebst, L., Levine, M., Bernasco, W., & Lindegaard, M. R. (2020). Would i be helped? Cross-national CCTV footage shows that intervention is the norm in public conflicts. *American Psychologist*, *75*(1), 66–75. https://doi.org/10.1037/amp0000469

Reid, T. (1764/1997). *An inquiry into the human mind on the principles of common sense*. Pennsylvania State University Press.

Romero, F. (2018). Who should do replication labor? *Advances in Methods and Practices in Psychological Science*, *1*(4), 516–537. https://doi.org/10.1177/2515245918803619

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass*, 14:e12633. https://doi.org/10.1111/phc3.12633

Rose, D., Buckwalter, W., & Nichols, S. (2017). Neuroscientific prediction and the intrusion of intuitive metaphysics. *Cognitive Science*, *41*, 482–502. https://doi.org/10.1111/cogs.12310

Rose, D., Buckwalter, W., & Turri, J. (2014). When words speak louder than actions: Delusion, belief, and the power of assertion. *Australasian Journal of Philosophy*, *92*(4), 683–700. https://doi.org/10.1080/00048402.2014.909859

Rose, D., Machery, E., Stich, S., Alai, M., Angelucci, A., Berniūnas, R., Buchtel, E. E., Chatterjee, A., Cheon, H., Cho, I.-R., Cohnitz, D., Cova, F., Dranseika, V., Lagos, Á. E., Ghadakpour, L., Grinberg, M., Hannikainen, I., Hashimoto, T., Horowitz, A., … Zhu, J. (2019). Nothing at stake in knowledge. *Noûs*, *53*(1), 224–247. https://doi.org/10.1111/nous.12211

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Russell, B. (2008 [1917]). *Mysticism and logic and other essays*. Project Gutenberg. https://www.gutenberg.org/ebooks/25447

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E.-J., & Aczel, B. (2022). A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, *9*(211997). https://doi.org/10.1098/rsos.211997

Scargle, D., Jeffrey. (2000). Publication bias: The "file-drawer" problem in scientific inference. *Journal of Scientific Exploration*, *14*(1), 91–106.

Schimmack, U. (2017). *Preliminary 2017 replicability rankings of 104 psychology journals: Vols. October 8, 2019* (October 8, 2019). https://replicationindex.com/2017/10/24/preliminary-2017-replicability-rankings-of-104-psychology-journals/

Schimmack, U. (2019). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science*, 1745691619863798. https://doi.org/10.1177/1745691619863798

Schwitzgebel, E., & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition*, *141*, 127–137. https://doi.org/10.1016/j.cognition.2015.04.015

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309–316. https://doi.org/10.1037/0033-2909.105.2.309

Selwood, A., Harris, C. B., Barnier, A. J., & Sutton, J. (2020). Effects of collaboration on the qualities of autobiographical recall in strangers, friends, and siblings: Both remembering partner and communication processes matter. *Memory*, *28*(3), 399–416. https://doi.org/10.1080/09658211.2020.1727521

Shewach, O. R., Sackett, P. R., & Quint, S. (2019). Stereotype threat effects in settings with features likely versus unlikely in operational test settings: A meta-analysis. *Journal of Applied Psychology*, *104*(12), 1514–1534. https://doi.org/10.1037/apl0000420

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(160384). https://doi.org/10.1098/rsos.160384

Stich, S. P., & Machery, E. (2022). Demographic differences in philosophical intuition: A reply to Joshua Knobe. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-021-00609-7

Strawson, G. (1986). *Freedom and belief*. Oxford University Press.

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

Strickland, B., & Suben, A. (2012). Experimenter philosophy: The problem of experimenter bias in experimental philosophy. *Review of Philosophy and Psychology*, *3*(3), 457–467. https://doi.org/10.1007/s13164-012-0100-9

Stroebe, W. (2019). What can we learn from many labs replications? *Basic and Applied Social Psychology*, *41*(2), 91–103. https://doi.org/10.1080/01973533.2019.1577736

Sytsma, J., & Livengood, J. (2011). A new perspective concerning experiments on semantic intuitions. *Australasian Journal of Philosophy*, *89*(2), 315–332. https://doi.org/10.1080/00048401003639832

Turri, J. (2014). The problem of ESEE knowledge. *Ergo*, *1*(4), 101–127. https://doi.org/10.3998/ergo.12405314.0001.004

Turri, J. (2017). Epistemic contextualism: An idle hypothesis. *Australasian Journal of Philosophy*, *95*(1), 141–156. https://doi.org/10.1080/00048402.2016.1153684

Turri, J., Buckwalter, W., & Blouw, P. (2015). Knowledge and luck. *Psychonomic Bulletin and Review*, *22*, 378–390. https://doi.org/10.3758/s13423-014-0683-5

Turri, J., Buckwalter, W., & Rose, D. (2016). Actionability judgments cause knowledge judgments. *Thought*, *5*(3), 212–222. https://doi.org/10.1002/tht3.213

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113

Van Inwagen, P. (2015). *Metaphysics*. Westview Press.

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Aert, R. C. M. van, & Assen, M. A. L. M. van. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832–1832. https://doi.org/10.3389/fpsyg.2016.01832

Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology*, *25*(6), 813–836. https://doi.org/10.1080/09515089.2011.631995

Wilholt, T. (2012). Epistemic trust in science. *The British Journal for the Philosophy of Science*, *64*(2), 233–253. https://doi.org/10.1093/bjps/axs007

Williams, B. A. O. (1981). *Moral luck*. Cambridge University Press.

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*, 16. https://doi.org/10.33735/phimisci.2022.9193