Philosophy of Science and The Replicability Crisis

Felipe Romero
University of Groningen[1]

Abstract. Replicability is widely taken to ground the epistemic authority of science. However, in recent years, important published findings in the social, behavioral, and biomedical sciences have failed to replicate, suggesting that these fields are facing a "replicability crisis." For philosophers, the crisis should not be taken as bad news but as an opportunity to do work on several fronts, including conceptual analysis, history and philosophy of science, research ethics, and social epistemology. This article introduces philosophers to these discussions. First, I discuss precedents and evidence for the crisis. Second, I discuss methodological, statistical, and social-structural factors that have contributed to the crisis. Third, I focus on the philosophical issues raised by the crisis. Finally, I discuss several solution proposals and highlight the gaps that philosophers could focus on.

(5600 words)

Introduction

Replicability is widely taken to ground the epistemic authority of science: we trust scientific findings because experiments repeated under the same conditions produce the same results. Or so one would expect. However, in recent years, important published findings in the social, behavioral, and biomedical sciences have failed to replicate (i.e., when independent researchers repeat the original experiment they do not obtain the original result.) The failure rates are alarming, and the growing consensus in the scientific community is that these fields are facing a "replicability crisis."

Why should we care? The replicability crisis undermines scientific credibility. This, of course, primarily affects scientists. They should clean up their acts and revise entire research programs to reinforce their shaky foundations. However, more generally, the crisis affects all consumers of science. We can justifiably worry that scientific testimony might lead us astray if many findings

[1] Department of Theoretical Philosophy, Faculty of Philosophy, University of Groningen; c.f.romero@rug.nl

that we trust unexpectedly fail to replicate later. And when we want to defend the epistemic value of science (e.g., against the increasing charges of partisanship in public and political discussions), it certainly does not help that the reliability of several scientific fields is doubtable. Additionally, as members of the public, the high replication failure rates are disappointing as they suggest that scientists are wasting taxpayer funds.

For philosophers, the replicability crisis also raises pressing issues. First, we need to address deceptively simple questions, such as "what is a replication?" Second, the crisis also raises questions about the nature of scientific error and scientific progress. While philosophers of science often stress the fallibility of science, they also expect science to be self-corrective. Nonetheless, the replicability crisis suggests that some portions of science may not be self-correcting, or, at least, not in the way in which philosophical theories would predict. In either case, we need to update our philosophical theories about error correction and scientific progress. Finally, the crisis also urges philosophers to engage in discussions to reform science. These discussions are happening in scientific venues, but philosophers' theoretical work (e.g., foundations of statistics) can contribute to them.

The purpose of this article is to introduce philosophers to the discussions about the replicability crisis. First, I introduce the replicability crisis, presenting important milestones and evidence that suggests that many fields are indeed in a crisis. Second, I discuss methodological, statistical, and social-structural factors that have contributed to the crisis. Third, I focus on the philosophical issues raised by the crisis. And finally, I discuss solution proposals emphasizing the gaps that philosophers could focus on, especially in the social epistemology of science.

1. What is the Replicability Crisis? History and Evidence

Philosophers (Popper, 1959/2002), methodologists (Fisher, 1926), and scientists (Heisenberg, 1975) take replicability to be the mark of scientific findings. As an often-cited quote by Popper observes, "non-replicable single occurrences are of no significance to science" (1959, p. 64). Recent discussions focus primarily on the notion of *direct replication*, which refers roughly to "repetition of an experimental procedure" (Schmidt, 2009, p. 91). Using this notion, we can state the following principle: Given an experiment *E* that produces some result *F*, *F* is a scientific finding

only if in principle a direct replication of *E* produces *F*. That is, if we repeated the experiment we should obtain the same result.

Strictly speaking, it is impossible to repeat an experimental procedure exactly. Hence, direct replication is more usefully understood as an experiment whose design is identical to an original experiment's design in all factors that are supposedly causally responsible for the effect. Consider the following example from Gneezy et al. (2014). The experiment *E* compares the likelihood of choosing to donate to a charity when the donor is informed that (a) the administrative costs to run the charity have already been covered or (b) that her contribution will cover such costs. *F* is the finding that donors are more likely to donate to a charity in the first situation. Imagine we want to replicate this finding directly (as Camerer et al., 2018, did). Changing the donation amount might make a difference; hence, the replication would not be direct, but whether we conduct the replication in a room with grey or white walls should be irrelevant.

A second notion that researchers often use is *conceptual replication*: "Repetition of a test of a hypothesis or a result of earlier research work with different methods" (Schmidt, 2009, p. 91). Conceptual replications are epistemically useful because they modify aspects of the original experimental design to test its generalizability to other contexts. For instance, a conceptual replication of Gneezy et al.'s experiment could further specify the goals of the charities in the vignettes, as these could influence the results as well. Additionally, methodologists distinguish replicability from a third notion: *reproducibility* (Peng, 2011; Patil et al., 2016). This notion means obtaining the same numerical results when repeating the analysis using the original data and the same computer code. Some studies do not pass this minimal standard.

Needless to say, these notions are controversial. Researchers disagree about how to best define them and the epistemic import of the practices that they denote (See Section 3 for further discussion). For now, these notions are useful to introduce four precedents of the replicability crisis:

- **Social priming controversy**. In the early 2010s, researchers reported direct replication failures of John Bargh's famous elderly-walking study (Bargh et al., 1996) in two (arguably better conducted) attempts (Pashler et al., 2011; Doyen et al., 2012). Before the failures, Bargh's finding had been positively cited for years, taught to psychology students, and it had inspired a big industry of "social priming" papers (e.g., many conceptual replications of Bargh's work).

Several of these findings have also failed to replicate directly (Harris, Coburn, Rohrer, & Pashler, 2013; Pashler, Coburn, & Harris, 2012; Shanks et al., 2013, Klein et al., 2014).

- **Daryl Bem's extrasensory perception studies**. Daryl Bem showed in nine experiments that people have ESP powers to perceive the future. His paper was published in a prestigious psychology journal (Bem, 2011). While the finding persuaded very few scientists, the controversy engendered mistrust in the ways psychologists conduct their experiments since Bem used procedures and statistical tools that many social psychologists use. (See Romero 2017, for discussion.)
- **Amgen and Bayer Healthcare reports**. Two often-cited papers reported that scientists from the biotech companies Amgen (Begley and Ellis, 2012) and Bayer Healthcare (Prinz et al., 2011) were only able to replicate a small fraction (11%~20%) of landmark findings in pre-clinical research (e.g., oncology), which suggested that replicability is a pervasive problem in biomedical research.
- **Studies on P-hacking and Questionable Research Practices**. Several studies (Ioannidis et al., 2008; Simmons et al., 2011; John et al., 2012; Ioannidis et al., 2014) showed how some practices that exploit the flexibility in data collection could lead to the production of false positives (see Section 2 for explanation). These studies suggested that the published record across several fields could be polluted with nonreplicable research.

While the precedents above suggested that there was something flawed in social and biomedical research, the more telling evidence for the crisis comes from multi-site projects that assess replicability systematically. In psychology, the Many Labs projects (Ebersole et al., 2016; Klein et al., 2014; Open Science Collaboration 2012) have studied a variety of findings and whether they replicate across multiple laboratories. Moreover, the Reproducibility Project (Open Science Collaboration, 2015), studied a random sample of published studies to estimate the replicability of psychology more generally. Similar projects have assessed the replicability of cancer research (Nosek & Errington, 2017), experimental economics (Camerer et al., 2016), and studies from the prominent journals *Nature* and *Science* (Camerer et al., 2018). These studies give us an unsettling perspective. The Reproducibility Project, in particular, suggests that only a third of findings in psychology replicate.

Now, it is worth noting that the concern about replicability in the social sciences is not new. What authors call the replicability crisis started around 2010, but researchers had been voicing concerns about replicability long before. As early as the late 1960s and early 1970s, authors worried about the lack of direct replications (Ahlgren, 1969; Smith, 1970). Also in the late 1970s, the journal *Replications in Social Psychology* was launched (Campbell and Jackson, 1979) to address the problem that replication research was hard to publish, but it went out of press after just three issues. Later in the 1990s, studies reported that editors and reviewers were biased against publishing replications (Neuliep & Crandall, 1990; Neuliep & Crandall, 1993). This history is instructive and triggers questions from the perspective of the history and philosophy of science. If researchers have neglected replication work systematically, isn't it unsurprising that many published findings do not replicate? Also, why hasn't the concern about replicability led to sustainable changes?

2. Causes of the Replicability Crisis

Most likely, the replicability crisis is the result of the interaction of multiple methodological, statistical, and sociological factors. (Although it is worth mentioning that authors disagree about how much each factor contributes.) Here I review the most discussed ones.

Arguably one of the strongest contributing factors to the replicability crisis is *publication bias*, i.e., using the outcome of a study (in particular, whether it succeeds supporting its hypothesis, and especially if the hypothesis is surprising) as the primary criterion for publication. For users of Null Hypothesis Significance Testing (NHST), as most fields affected by the crisis, publication bias results from making statistical significance a necessary condition for publication. This leads to what Rosenthal in the late 1970s labeled "the file-drawer problem" (Rosenthal, 1979). By chance, a false hypothesis is expected to be statistically significant 5% of the time (following the standard convention in NHST). If journals only publish statistically significant results, then they contain the 5% of the studies that show erroneous successes (false positives) while the other 95% of the studies (true negatives) remain in the researchers' file drawers. This produces a misleading literature and biases meta-analytic estimates. Publication bias is more worrisome when we consider that only a fraction of all the hypotheses that scientists test are true. In such a case, it is possible that most published findings are false (Ioannidis, 2005). Recently, methodologists have developed

techniques to identify publication bias (Simonsohn, Nelson, & Simons, 2014; van Aert, Wicherts, & van Assen, 2016).

Publication bias fuels a second contributing factor to the replicability crisis, namely, *Questionable Research Practices* (QRPs). Since statistical significance determines publication, scientists have incentives to deviate (sometimes even unconsciously) to achieve it. For instance, scientists anonymously admit that they engage in a host of QRPs (John et al., 2012), such as reporting only studies that worked. A particularly pernicious practice is *p-hacking*, that is, exploiting the flexibility of data collection to obtain statistical significance. This includes, e.g., collecting more data or excluding data until you get your desired results. In an important computer simulation study, Simmons et al. (2011) show that a combination of p-hacking techniques can increase the false positive rate to 61%. QRPs and p-hacking are troublesome because (1) unlike clear instances of fraud, they are widespread, and (2) motivated reasoning can lead researchers to justify them (e.g., "I think that person did not quite understand the instructions of the experiment, so I should exclude her data.")

Also related to publication bias, the *proliferation of conceptual replications* is a third factor that contributes to the replicability crisis. As discussed by Pashler and Harris (2012), the problem of conceptual replications lies in their interaction with publication bias. Suppose a scientist conducts a series of experiments to test a false theory T. Suppose he fails in all but one of his attempts; the only one that gets published. Then, a second scientist gets interested in the publication. She tries to test T in modified conditions in a series of conceptual replications, without replicating the original conditions. Again, she succeeds in only one of her attempts, which is the only one published. In this process, none of the replication failures gets published given the file-drawer problem. But still, after some time, the literature will contain a diverse set of studies that suggest that T is a robust theory. In short, the proliferation of conceptual replications might misleadingly support theories.

A fourth contributing factor to the replicability crisis is Null Hypothesis Significance Testing itself. The argument can take two forms. On the one hand, scientists' literacy on NHST is low. Already before the replicability crisis, authors argued that practicing scientists misinterpret p-values (Cohen, 1990), consistently misunderstand the inferential logic of the method (Fidler 2006), and confuse statistical significance with scientific import (Ziliak & McCloskey, 2008). Moreover, recently, the American Statistical Association explicitly listed the misunderstanding of NHST as a

cause of the crisis (Wasserstein & Lazar, 2016). On the other hand, there are concerns about the *limitations* of NHST. Importantly, in NHST, non-statistically significant results are typically inconclusive so researchers cannot accept a null hypothesis (but see Machery, 2012 and Lakens, Scheel, & Isager, 2018). And if we cannot accept a null hypothesis, then it is harder to evaluate and publish failed replication attempts.

A fifth and arguably more fundamental factor that contributes to the replicability crisis is the reward system of science. A central component of the reward system of science is the priority rule (Merton, 1957), i.e., the practice of rewarding only the first scientist that makes a discovery. This reward system discourages replication (Romero, 2017). The argument concerns the interaction between the priority rule and the peer-review system. In present-day science, scientists establish priority over a finding via peer-reviewed publication. However, since peer-review is insufficient to determine whether a finding replicates, many findings are rewarded with publication regardless of their replicability. The reward system also contributes to the production of non-replicable research by exerting high career pressures on researchers. They need to fill their CVs with exciting, positive results to sustain and advance in their careers. This perverse incentive explains why many of them fall prey to QRPs, confirmation biases (Nuzzo, 2015), and posthoc hypothesizing (Kerr, 1996; Bones, 2012), leading to non-replicable research.

3. Philosophical Issues Raised by the Replicability Crisis

Psychologists acknowledge the need for philosophical work in the context of the replicability crisis: they are publishing a large number of papers with conceptual work inspired by the crisis. Some authors voice the need for philosophy explicitly (Spellman, 2015, p.894). Philosophers, with few notable exceptions, are only recently joining these discussions. In this section, I review some of the more salient philosophical issues raised by the crisis and point out open research avenues.

The first set of philosophical issues triggered by the crisis concerns the very definition of replication. What is a replication? Methodologists and practicing scientists often use the notions of direct (i.e., "repetition of an experimental procedure", Schmidt, 2009, p. 91) and conceptual ("repetition of a test of a hypothesis or a result of earlier research work with different methods", Schmidt, 2009, p. 91) replication. Philosophers have made similar distinctions, albeit using different terminology (Cartwright, 1991; Radder, 1996). However, both notions are vague and

require further specification. While the notion of direct replication is intuitive, strictly speaking, no experiment can repeat the original study because there are always unavoidable changes in the setting, even if they are small (e.g., changes in time, weather, location, and participants.) One amendment, as suggested above, is to reserve the term "direct replication" for experiments whose design is identical to an original experiment's design in all factors that are supposedly causally responsible for the effect. The notion of conceptual replication is even more vague. This notion denotes the practice of modifying an original experimental design to evaluate a finding's generalizability across laboratories, measurements, and contexts. While this practice is fairly common, as researchers change an experiment's design, the resulting designs can be very different. These differences can lead researchers to disagree about what hypothesis the experiments are actually testing. Hence, labeling these experiments as "replications" can be controversial.

Authors have attempted to refine the definitions of replication to overcome the problems of the direct/conceptual dichotomy. One approach is to view the difference between the original experiment and replication as a matter of degree. The challenge is then to specify the possible and acceptable ways in which replications can differ. For instance, Brandt et al. (2014) suggest the notion of "close" replication. For them, the goal should be to make replications as close as possible to the original while acknowledging the inevitable differences. Similarly, LeBel et al. (2018) identify a replication continuum of five types of replications that are classified according to their relative methodological similarity to the original study. And Machery (2019a) argues that the direct/conceptual distinction is confused and defines replications as experiments that can resample several experimental components.

Having the right definition of replication is not only theoretically important but also practically pressing. Declaring that a finding fails to replicate depends on whether the replication attempt counts as a replication or not. In fact, the reaction of some scientists whose work fails to replicate is to emphasize that the replication attempts introduce substantive variations which explain the failures and list a number of conceptual replications that support the underlying hypothesis (for examples of this response, see Carney et al., 2015 and Schnall, 2014). The implicature in these responses is that the failed direct replication attempts are not genuine replications and the successful conceptual replications are.

The definitional questions trigger closely related epistemological questions. What is the epistemic function of replication? How essential are replications to further the epistemic goals of science? An immediate answer is that replications (i.e., direct or close replications) evaluate the reliability of findings (Machery 2019a). So understood, conducting replications serves a crucial epistemic goal. But some authors disagree. For instance, Stroebe and Strack (2018) argue that direct replications are uninformative because they cannot be exact and suggest to focus on conceptual replications instead. Similarly, Leonelli (2018) argues that in some cases, the validation of results does not require direct/close replications, and non-replicable research often has epistemic value. And Feest (2019) also argues that replication is only a very small part of what is necessary to improve psychological science, and hence, the concerns about replicability are overblown. These remarks urge researchers to reconsider their focus on replication efforts.

Another pressing set of philosophical questions triggered by the replicability crisis concerns the topic of scientific self-correction. For an important tradition in philosophy, science has an epistemically privileged position not because it gives us truth right away but because in the long-run it corrects its errors (Peirce, 1901/1958; Reichenbach, 1938). Authors call this idea the self-corrective thesis (Laudan, 1980; Mayo, 2005).

> **(SCT)** In the long run, the scientific method will refute false theories and find closer approximations to true theories.

In the context of modern science, we can refine SCT to capture the most straightforward mechanism of scientific self-correction, which involves replication, statistical inference, and meta-analysis.

> **(SCT*)** Given a series of replications of an experiment, the meta-analytical aggregation of their effect sizes will converge on the true effect size (with a narrow confidence interval) as the length of the series of replications increases.

SCT* is theoretically plausible but its truth depends on the social structural conditions that implement it. First, most findings are never subjected to even one replication attempt (Makel et al., 2012). It is true that scientists have recently identified particular findings that do not replicate, but this is a tiny step in the direction of self-correction. If we trust the estimates of low replicability, these failures could be the tip of the iceberg, and the false positives under the surface may never be corrected. Second, the social structural conditions in the fields affected by the crisis (which

involve publication bias, confirmation bias, and limited resources) make the thesis false (Romero, 2016). Now, the falsity of SCT* does not entail that SCT is false but requires us to specify what other mechanisms could make SCT true.

We can see the concern about SCT as an instance of a broader tension between the theory and practice of science. The replicability crisis reveals a gap between our image of science, which includes the ideal of self-correction via replication, and the reality (Longino, 2015). We can view this gap in several ways. One possibility is that the replicability crisis proves that the ideal is normatively inadequate (i.e., cannot implies not-ought). Hence, we have to change the ideal to close the gap, and this project requires philosophical work. Another possibility is that the ideal is adequate, and the gap is an implementation failure that results from bad scientists not doing their job. In this view, the gap is less philosophically significant and more a problem for science policymakers. In favor of the first possibility, however, it is worth stressing that many scientists succumb to practices that lead to non-replicable research. That is, the gap is not due to a few bad apples but to systemic problems. This assessment invites social epistemological work.

The replicability crisis also raises questions about confirmation, specifically regarding the variety of evidence thesis (VET). This thesis states that ceteris paribus varied evidence (e.g., distinct experiments pointing to the same hypothesis from multiple angles) has higher confirmatory power than less varied evidence. (This idea is also discussed in philosophy under the labels of "robustness analysis" and "triangulation.") VET has intuitive appeal and has been favorably appraised by philosophers (Wimsatt 1981; see Landes, 2018 for discussion.) Take, for instance, the case for climate change, which we take to be robust as it incorporates evidence from a variety of different disciplines. Nonetheless, VET is not uncontroversial (Stegenga, 2009). In the context of the crisis, the virtues of VET need to be qualified, given the concern that conceptual replications have contributed to the problem (see Section 2). Since the 1990s, in line with VET, a model paper in psychology contains a series of distinct experiments testing the same hypothesis with conceptual replications. While such a paper allegedly gives a robust understanding of the phenomenon, the conceptual replications in many cases have been conducted under the wrong conditions (e.g., confirmation bias, publication bias, and low statistical power), and are therefore not trustworthy (Schimmack, 2012). In these cases, having more direct replications (i.e., less varied evidence) could even be more epistemically desirable. Thus, the replicability crisis requires us to evaluate VET from a practical perspective and determine when conceptual replications confirm or mislead.

Another concern that the replicability crisis raises for philosophers has to do with epistemic trust. Science requires epistemic trust to be efficient (Wilholt, 2013). But how much should you trust? Scientists cannot check all the findings they rely on. If they did, science would be at best inefficient. However, in light of the replicability crisis scientists cannot be content trusting the findings of their colleagues only because they are published. Epistemic trust can also lead consumers of non-replicable research from other disciplines astray. For example, empirically-informed philosophers, and specifically moral psychologists, have relied heavily on findings from social psychology. They also need to clean up their act. (See Machery & Doris, 2017, for suggestions on how to do this.)

While the issues above are primarily epistemological, the replicability crisis also raises ethical questions that philosophers have yet to study. A first issue concerns *research integrity* to facilitate replicability. A second issue concerns the *ethics of replication* itself. Since the first replication failures of social psychological effects in the early 2010s, the psychological community has witnessed a series of unfortunate exchanges. Original researchers have questioned the competence of replicators and even accused them of ill-intent and bullying (Yong, 2012; Meyer & Chabris 2014; Bohannon, 2014). What should we make of these battles? While the scientific community has the right to criticize any published finding, replication failures can impact on original researchers' careers dramatically (e.g., affecting hiring and promotion.) Replicators can make mistakes too. In recent years, there has been a growing movement of scientists focused on checking the work of their colleagues. While the crisis epistemically justifies their motivation, it is also fair to ask, who checks the checkers?

4. What to do?

The big remaining question is normative: what should we do? Since the crisis is likely the result of multiple contributing factors, there is a big market of proposals. I classify them in three camps: *statistical reforms*, *methodological reforms,* and *social reforms*. I use this classification primarily to facilitate discussion. Indeed, there are few strict reformists of each camp. Most authors agree that science needs more than one kind of reform. Nonetheless, authors also tend to emphasize the benefits of particular interventions (in particular, the statistical reformists). I discuss some of the most salient proposals from each camp.

4.1. Statistical Reforms

Statistical reformists are of two kinds. The first kind advocates for replacing frequentist statistics (in particular, NHST). One alternative is to completely get rid of NHST and use descriptive statistics instead (Trafimov & Marks, 2015). A more prominent approach is Bayesian inference (Bernardo & Smith 1994; Rouder et al. 2009; Lee & Wagenmakers 2013). The argument for Bayesian inference is foundational. The Bayesian researcher needs to be explicit about several assumptions in her tests–assumptions that remain under the hood of NHST inference (Romeijn, 2014; Sprenger, 2016). Additionally, Bayesian inference with Bayes factors (the most popular measure of evidence for Bayesian inference in psychology) gives the researcher a straightforward procedure to infer a null hypothesis. This is a great advantage when dealing with replication failures. In practice, however, authors disagree about how to specify the necessary assumptions to conduct Bayes factor analysis.

The second kind of statistical reformist does not want to eliminate frequentist statistics but change the way we do frequentist statistics. There are philosophical motivations for this sort of reform. Long-run error control is a valuable goal of statistical inference, which is not clearly met outside frequentism. Hence, rather than replacing frequentist statistics, one may argue that we need to improve the way we use it (Mayo, 2018). The frequentist scientists have had tools in addition to p-values to make inferences that practitioners could incorporate. For instance, equivalence tests allow researchers to test for the absence of effects (Lakens, Scheel, & Isager, 2018). Another possibility is to move away from the dichotomous inferential approach of NHST and focus on estimating effect sizes and confidence intervals (Fidler, 2007; Cumming, 2012; Cumming, 2014).

There are also practical motivations to preserve frequentist statistics and in particular NHST. For instance, Benjamin et al. (2018) in a 72 authors paper advocate for changing the p-value threshold from the conventional $p<0.05$ to the stricter $p<0.005$. While the authors acknowledge the problems of NHST, they argue that such a change would solve many of the problems that lead to low replicability (e.g., by making p-hacking and QRPs harder to work) and would be easy to implement. In response, other 88 authors argue for a more critical approach in which authors should be required to specify and justify the significance level that their project needs (Lakens et al., 2018).

While philosophers are less invested in developing new statistical tools, they can contribute to these discussions at least in three ways: (1) evaluating the arguments and tradeoffs involved in implementing statistical reforms (Machery, 2019b); (2) making foundational debates about statistical inference relevant and accessible to practitioners; and (3) studying how inference methods behave in different contexts, e.g., by using computer simulations (Romero, 2016, Bruner & Holman, 2019, and Romero and Sprenger, manuscript).

4.2. Methodological Reforms

The methodological reformist proposes to improve scientific practices more generally by going beyond mere statistics. One type of reform is explicitly *anti-statistical*. For instance, McShane et al. (2019) argue to make publication decisions considering statistical outcomes (e.g., p-values, confidence intervals, and Bayes Factors) as just another piece of information among other factors such as "related prior evidence, plausibility of mechanism, study design and data quality, real-world costs and benefits, and novelty of finding" (p. 235). In practice, however, it would be hard to implement alternatives like this because editors and reviewers are used to relying on statistical thresholds as heuristics to make publication decisions.

The second type of methodological reform recommends making the scientific process more transparent. A popular movement with this aim is *open science*. The rationale of open science practices is to increase transparency by asking researchers to share a variety of products from their work, ranging from experimental designs to software and raw data. Open science is epistemically desirable (but see Levin & Leonelli, 2016). Specifically, in the context of the crisis, open science practices have the potential to increase replicability as they greatly facilitate replication work by independent researchers.

The open science movement has also defended *pre-registration* enthusiastically. That is, uploading a timestamped uneditable research plan to a public archive. A pre-registration states the hypotheses to be tested, target sample sizes, and so on. Pre-registration greatly constrains the researcher degrees of freedom that make QRPs and p-hacking work. When authors submit their work to a journal, reviewers and editors can verify whether the authors did what they planned.

While pre-registration increases transparency, we should not overstate its usefulness. First, pre-registration does not fully counter publication bias as it does not guarantee that findings will be

reported (Chen et al., 2016). Second, pre-registration cannot be straightforwardly implemented in some research domains (Tackett et al., 2017). Two refinements on pre-registration are the Registered Reports (Chambers, 2013) and Registered Reproducibility Reports (Simons et al., 2014) publication models. In these models, scientists submit a research proposal to a journal before data collection, which is evaluated based on its methodological merits. The journal can give the proposal an in-principle acceptance (IPA), which means that the paper will be published regardless of its outcome (see Romero, 2018, for discussion).

Various authors have proposed changing publication practices to address the problem that replications are not rewarded. As discussed in Section 1, having dedicated outlets for replication work has not worked in the past. The reason is likely that having replication work published in secondary venues gives the impression that such a work is not very important, and hence researchers would still relegate it. Instead, a more promising approach is opening the doors of prestigious journals for replication work (Cooper, 2016; Simons, Holcombe, & Spellman, 2014; Vazire, 2015).

4.3. Social Reforms

The social reformist argues that changes in statistics and methodology are insufficient to address the replicability crisis because they treat the symptoms and not the disease, namely the defective social structures of contemporary science. For the social reformist, it is too optimistic to expect scientists to follow good practices (in particular, to do replication work) if the right incentives are not in place. This is because science today is a professionalized activity. As such, scientists are constrained not only by the ethos of science but also by more mundane and arguably more forceful pressures, such as the requirement to produce many novel findings to have a career and continue playing the game.

Social reforms attempt to align career incentives with statistical and methodological expectations. In particular, to incentivize replication work, multiple parties should intervene. Funding agencies can allocate funding specifically to replication projects (see Netherlands Organisation for Scientific Research, 2016, for an example). Universities and departments can create positions in which replication work is part of the responsibility of the researcher, and they can adapt promotion criteria according to quality metrics rather than raw publication numbers

(Schönbrodt, Heene, Maier, & Zehetleitner, 2015). Such interventions would create conditions where researchers do not perceive replication as second-class work.

Further questions that the social reformist asks concern the adequate design of epistemic institutions: What is the best way to divide cognitive labor to ensure that science produces novel findings but also replicable results? What are the different tradeoffs in terms of speed and reliability if we incorporate replication work as an essential part of the research process? Should all scientists in a community engage in replication work or only a selected group? Some authors answer these questions proposing different institutional arrangements (see Romero, 2018, for discussion) and this is an area ripe for social epistemological investigation.

Conclusion

In this paper, I have reviewed core issues in discussions around the replicability crisis, including its history, causes, philosophical assessments, and proposed solutions. Many normative discussions about replicability focus on technical problems about statistical inference and experimental design. Philosophers with interest on the foundations of statistical inference and confirmation theory can play a more active role in them. But the replicability crisis is not exclusively (and not primarily) a statistical problem. As I have reviewed, we still need to clarify concepts about replication, understand how different practices impact on low replicability and study how to intervene in the social structure of science. In these respects, the crisis demands work from the perspectives of the history and philosophy of science, social epistemology, and research ethics. That is, for philosophers, the crisis should not be taken as bad news but as an opportunity to update our theories and make them relevant to practice.

Acknowledgments

I am grateful to Mike Dacey, Teresa Ai, the editor, and one anonymous reviewer for their useful comments on previous drafts.

References

Ahlgren A, (1968). A modest proposal for encouraging replication. *American Psychologist*, 24, 471.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230–244. doi:10.1037/0022-3514.71.2.230

Begley, C. G., & Ellis, L. M. (2012). Drug Development: Raise standards for preclinical Cancer research. Nature, 483, 531-533.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100,* 407–425. doi:10.1037/a0021524

Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian Theory*. New York: Wiley.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., …, and Johnson, V. E. (2018). Redefine statistical significance. Nature Human Behaviour 2, 6–10.

Bohannon, J. (2014). Replication effort provokes praise—and 'bullying' charges. *Science, 344*, 788–789. doi:10.1126/science.344.6186.788

Bones, A. K. (2012). We knew the future all along: Scientific hypothesizing is much more accurate than other forms of precognition —a satire in one part. *Perspectives on Psychological Science, 7*, 307–309. doi:10.1177/1745691612441216

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

Bruner, J.P, & Holman, B., (2019). Self-correction in science: Meta-analysis, bias and social structure. *Studies in History and Philosophy of Science Part A.*

Camerer, C. F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, … , and Hang Wu (2016). Evaluating Replicability of Laboratory Experiments in Economics. *Science*. https://doi.org/10.1126/science.aaf0918.

Campbell, K. E., & Jackson, T. T. (1979). The role and need for replication research in social psychology. *Replications in Social Psychology, 1*(1), 3–14.

Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science, 21*, 1363–1368.

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2015). Review and summary of research on the embodied effects of expansive (vs. contractive) nonverbal displays. *Psychological Science, 26*(5), 657–663.

Cartwright, N. (1991) Replicability, Reproducibility and Robustness: Comments on Harry Collins, *History of Political Economy*, *23*(1): 143–155.

Chambers, C. D. (2013). *Registered Reports*: A new publishing initiative at *Cortex*. *Cortex, 49*, 609–610. doi:10.1016/j.cortex.2012.12.016

Chen, R., Desai, N. R., Ross, J. S., Zhang, W., Chau, K. H., Wayda, B., . . . Krumholz, H. M. (2016). Publication and reporting of clinical trial results: Cross sectional analysis across academic medical centers. *BMJ, 352*, i637. doi:10.1136/bmj.i637

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304–1312.

Cooper, M. L. (2016). Editorial. *Journal of Personality and Social Psychology, 110*, 431–434. doi:10.1037/pspp0000033

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M. and Cushman, F., … 2018. Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 1–36.

Cumming, G. (2012). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. Multivariate applications book series. Routledge.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science 25*, 7–29.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLOS ONE, 7*(1), Article e29081. doi:10.1371/journal.pone.0029081

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . 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.

Feest, Uljana (2019). Why replication is overrated, *Philosophy of Science*.

Fidler, F. (2006). Should Psychology Abandon p-values and Teach CIs Instead? Evidence-Based Reforms in Statistics Education. *In Proceedings of the 7th International Conference on Teaching Statistics.*

Fidler, F. (2007). *From Statistical Significance to Effect Estimation: Statistical Reform in Psychology, Medicine and Ecology*. https://doi.org/10.1080/13545700701881096.

Fisher, R. A. (1926). The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture*, *33*, 503–13.

Gneezy, U., Keenan, E.A., and Gneezy, A. (2014). Avoiding Ovehead Aversion in Charity, *Science*, 346 (6209), 632–35.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLoS ONE*, 8, e72467.

Heisenberg, W. (1975). The Great Tradition: End of an Epoch?, *Encounter*, 44 (3), 52–58.

Horwich, P. (1982). *Probability and evidence*. Cambridge: Cambridge University Press

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*, 640–648. doi:10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A., Munafò, 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*, 235–241. doi:10.1016/j.tics.2014.02.010

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–532. doi:10.1177/0956797611430953

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217. doi:10.1207/s15327957pspr0203_4

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahnik, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology, 45*, 142–152. doi:10.1027/1864-9335/a000178

Koole, S. L., & Lakens, D. (2012). Rewarding replications. *Perspectives on Psychological Science, 7,* 608–614. doi:10.1177/1745691612462586

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., … , and Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour,* 2, 168–171.

Landes, J. (2018). Variety of Evidence, *Erkenntnis*.

Laudan, L. (1981). *Science and Hypothesis: Historical Essays on Scientific Methodology*. Springer Netherlands.

Lee, M. D. & Wagenmakers, E-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge: Cambridge University Press.

Leonelli, S. (2018) Re-Thinking Reproducibility as a Criterion for Research Quality. [Preprint]: http://philsci-archive.pitt.edu/14352/1/Reproducibility_2018_SL.pdf

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. Journal of Personality and Social Psychology, 113, 254–261. doi:10.1037/pspi0000106

Levin, N., & Leonelli, S. (2016). How Does One "Open" Science? Questions of Value in Biological Research. *Science, Technology, & Human Values*, 42(2), 280-305

Longino, H. (2015). The social dimensions of scientific knowledge. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2015 ed.)

Machery, E. (2012). Power and Negative Results. *Philosophy of Science*, 79 (5): 808–20.

Machery, E. (2019a). What is a Replication? [Preprint]

Machery, E. (2019b). The Alpha War. *Review of Philosophy and Psychology*.

Machery, E., & Doris, J. M. 2017. An open letter to our students: Doing interdisciplinary moral psychology. In B. G. Voyer and T. Tarantola (Eds), *Moral Psychology: A multidisciplinary guide*, 119-143. Springer.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019) Abandon Statistical Significance, *The American Statistician*, 73:sup1, 235–245, doi:10.1080/00031305.2018.1527253

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*, 537–542. doi:10.1177/1745691612460688

Mayo, D. (2005). Peircean induction and the error-correcting thesis. *Transactions of the Charles S. Peirce Society*, 41, 299-319.

Mayo, D. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Science Wars*. Cambridge: Cambridge University Press.

Merton, R. K. (1957). Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review, 22*, 635–659. doi:10.2307/2089193

Meyer, M. N., & Chabris, C. (2014, July 31). Why psychologists' food fight matters. *Slate*. Retrieved from http://www.slate.com/articles/health_and_science/science/2014/07/replication_controversy_in_psychology_bullying_file_drawer_effect_blog_posts.html

Netherlands Organisation for Scientific Research. (2016, July 16). NWO makes 3 million available for Replication Studies pilot. Retrieved from https://www.nwo.nl/en/news-and-events/news/2016/nwo-makes-3-million-available-for-replication-studies-pilot.html

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior & Personality, 5*(4), 85-90.

Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior & Personality, 8*(6), 21-29.

Nosek, B. A & Errington, T. M. (2017). Reproducibility in cancer biology: Making sense of replications. *eLife 6*, e23383.

Nuzzo, R. (2015). How scientists fool themselves – and how they can stop. *Nature, 526*, 182–185. doi:10.1038/526182a

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. doi:1126/science.aac4716

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social Distance? Failure to replicate effects on social and food judgments. *PLoS ONE*, 7, e42510.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536. doi:10.1177/1745691612463401

Pashler, H., Harris, C. R., & Coburn, N. (2011). Elderly-related words prime slow walking. *PsychFileDrawer*. Retrieved from http://www.PsychFileDrawer.org/replication.php?attempt=MTU%3D

Patil P., Peng R. D., Leek J. T. (2016). A statistical definition for reproducibility and replicability. *bioRxiv*. doi: 10.1101/066803

Peirce, C. S. (1958). The logic of drawing history from ancient documents. In A. W. Burks (Ed.), *The collected papers of Charles Sanders Peirce* (Vol. IV, pp. 89–107). Cambridge, MA: Belknap Press. (Original work published 1901)

Peng R. D. (2011). Reproducible research in computational science. *Science*, 334, 1226–1227. doi: 10.1126/science.1213847

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or Not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10, 712.

Popper, K. R. (1959/2002). *The Logic of Scientific Discovery*. Classics Series. London: Routledge.

Radder, H. (1996). *In And About The World: Philosophical Studies Of Science And Technology*, Albany, NY: State University of New York Press.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science, 26*, 653–656. doi:10.1177/0956797614553946

Reichenbach, H. (1938). *Experience and prediction*. Chicago, IL: University of Chicago Press.

Rouder, J. N., Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson (2009). Bayesian *t* Tests for Accepting and Rejecting the Null Hypothesis. *Psychonomic Bulletin & Review 16*, 225–237.

Romeijn, J-W. (2014). Philosophy of Statistics. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/sum2018/entries/statistics/.

Romero, F. (2016). Can the behavioral sciences self-correct? A social epistemic study. *Studies in History and Philosophy of Science Part A, 60*, 55–69. doi:10.1016/j.shpsa.2016.10.002

Romero, F. (2017) Novelty vs. Replicability: Virtues and Vices in the Reward System of Science, *Philosophy of Science*, 84(5), 1031-1043.

Romero, F. (2018) Who Should Do Replication Labor?, *Advances in Methods and Practices in Psychological Science*, 1(4), 516-537.

Romero, F., & Sprenger, J. (2019). Scientific Self-Correction: The Bayesian Way. Retrieved from: https://psyarxiv.com/daw3q/

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18*, 682–689. doi:10.3758/s13423-011-0088-7

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*(4), 551-566.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13, 90-100.

Schnall, S. (2014) Further thoughts on replications, ceiling effects and bullying. Retrieved from: http://www.psychol.cam.ac.uk/cece/blog

Schönbrodt, F., Heene, M., Maier, M., & Zehetleitner, M. (2015). *The replication-/credibility-crisis in psychology: Consequences at LMU?* Retrieved from https://osf.io/nptd9/

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., . (2013). Priming intelligent Behavior: An elusive phenomenon. *PLoS ONE*, 8, e56515.

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*, 1359–1366. doi:10.1177/0956797611417632

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to Registered Replication Reports at *Perspectives on Psychological Science*. *Perspectives on Psychological Science, 9*, 552–555. doi:10.1177/1745691614543974

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*, 534–547. doi:10.1037/a0033242

Smith, N. C. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist, 25*, 970–975. doi:10.1037/h0029774

Spellman, B. A. (2015). A short (personal) future history of Revolution 2.0. *Perspectives on Psychological Science, 10*, 886–899. doi:10.1177/1745691615609918

Sprenger, J. (2016). Bayesianism vs. Frequentism in Statistical Inference. In *The Oxford Handbook of Probability and Philosophy*, pp. 185–209. Oxford University Press UK.

Stegenga, J. (2009). Robustness, Discordance, and Relevance. *Philosophy of Science, 75*(5), 650–661.

Strobe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*, 59–71.

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., . . . Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science, 12*, 742–756. doi:10.1177/1745691617690042

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*, 1–2. doi:10.1080/01973533.2015.1012991

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on *p* values: Reservations and recommendations for applying *p*-uniform and *p*-curve. *Perspectives on Psychological Science, 11*, 713–729. doi:10.1177/1745691616650874

Vazire, S. (2015). Editorial. *Social Psychological & Personality Science,7,* 3–7. doi:10.1177/1948550615603955

Wasserstein, R. L., and Nicole A. Lazar. 2016. The ASA's Statement on P-Values: Context, Process, and Purpose. *The American Statistician*, 70 (2), 129–33. http://amstat.tandfonline.com/doi/abs/10.1080/ 00031305.2016.1154108.

Wilholt, T. (2013). Epistemic Trust in Science, *British Journal for the Philosophy of Science*, 24(2), 233–253.

Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. In M. Brewer & B. Collins (Eds.), Scientific inquiry and the social sciences (pp. 124–163). San Francisco: Jossey-Bass.

Yong, E. (2012, March 10). A failed replication attempt draws a scathing personal attack from a psychology professor [Web log post]. *Discover Magazine Blog*. Retrieved from http://blogs.discovermagazine.com/notrocketscience/2012/03/10/failed-replication-bargh-psychology-study-doyen/