Incentivizing Replication is Insufficient to Safeguard Default Trust

Hugh Desmond

Forthcoming in *Philosophy of Science*

Abstract

Philosophers of science and meta-scientists alike now typically model scientists' behavior as driven by credit maximization. In this paper I argue that this modeling assumption cannot account for how scientists have a default level of trust in each other's assertions. The normative implication of this is that science policy should not only focus on incentive reform.

1. Introduction

When thinking about the social structures of science, philosophers of science and meta-scientists have for some time now predominantly adopted an 'economic approach', where scientists are modeled as credit-maximizing agents responding to incentives such as promotion, funding, or publication criteria (Kitcher 1990; Strevens 2006; Higginson and Munafò 2016; Smaldino and McElreath 2016; Heesen 2018; Holman and Bruner 2017; O'Connor 2019).

Yet in applied ethics, sociology of science, and to a large extent actual science policy making, an 'ethical approach' informs research on social structures of science: scientists are predominantly understood to be agents concerned with ideals such as honesty, respect, or reliability, and are capable of acting contrary to incentive structures (Carvalho 2017; Desmond 2019; ESF-ALLEA 2017; Forsberg et al. 2018; Godecharle et al. 2013).

Philosophers of science and meta-scientists do not openly dismiss the ethical approach. Yet, it does often not seem so clear what precisely, if anything, the ethical approach brings to the explanatory table that cannot be covered by the economic approach. For instance, concern with honesty could be explained as minimizing the expected penalties (negative credit) following a strategy of dishonesty; concern for reliability could be explained as maximizing replicable studies, which can be modeled as having a higher pay-off than non-replicable studies (as in Heesen 2018).

What is it, if anything, that prevents one from taking a cynical stance on the ideals of individual scientists, i.e., that they are mere window-dressing, ineffectual against the brutal reality of credit-maximization? Indeed, the view that scientists *should* be credit-maximizers, in interests of scientific progress, seems viable (Kitcher 1990).

The importance of this question extends to science policy, which often is implicitly informed by an economic approach. Consider for instance how, in order to increase individual scientists' honesty, some have called for increased penalties for fraud, even to the extent of making criminal prosecutions scientific misconduct more widespread (see e.g. Collier 2015). Similarly, it is proposed that individuals' concern for reliability can be improved by incentivizing replication research, for instance by giving funding and "badges" to scientists doing replication studies (Munafò et al. 2017), or by introducing "Replication Awards" (Gorgolewski et al. 2018).

This paper will suggest that the explanatory limits of the economic approach are reached concerning the phenomenon of *default trust between scientists*. This will be defined in detail later on (section 2), but the core idea is that scientists tend to believe that their colleagues are telling the truth – or are at least attempting to do so. This default trust in each others' assertions underlies many core scientific behaviors – I consider peer review and collaboration as illustrative examples – and thus may be considered integral to scientific research.

Given widespread problems with reproducibility and replicability[1] (Baker 2016) – or at least, the perception of such problems (Fanelli 2018) – such default trust can be said to be under pressure. I then consider one of the most important proposed policy reforms to the credit-based incentive structure of science: incentivizing replication research. Replication research basically acts to disincentivize low-credence assertions. Can it safeguard default trust? Using an expanded version of Heesen's model of when replicable (or trustworthy) assertions maximize credit (section 3), I will then show that, no matter how much replication research is incentivized, default trust cannot be justified in a culture of credit-maximization (section 4).

The upshot is that, on the descriptive side, the economic model cannot account for an important explanandum concerning scientific practice (i.e., default trust). On the normative side, it means that in a culture of extreme credit-maximization, default trust between scientists would ultimately be eroded, and replaced with a 'default lack of trust' (or default distrust), and this would be detrimental to science.

[1] I will adopt the National Academy of Science's definition of reproducibility and replicability (basically: obtaining consistent results by redoing the same analysis on the same data, versus confirming a hypothesis with different data and/or methods). See (NAS 2019, p. 46).

2. Default Trust in Scientific Research

In the literature on trust (for summary, see Hawley 2012), trust in an assertion *p* is typically understood to depend on the asserting agent's knowledge that *p* and honesty. If the agent either lacks knowledge, or is dishonest, trust would be inappropriate. For the particular issue of the trustworthiness of scientists' assertions (i.e., observations, hypotheses, theories), whether scientists operate with a knowledge norm of assertion (Williamson 2002), or merely posit empirically successful hypotheses is, of course, a vexed issue concerning epistemology and scientific realism that I would like to sidestep here. Instead I will analyze a scientist's assertion φ as type of action, where the trustworthiness of an action depends on the agent's *competence* to carry out φ successfully and *intention* to carry out φ (see Hawley 2012). For example, if a climate scientist tells me that humans are responsible for global warming, I will trust that assertion when I believe that the climate scientist has the right type of competence (understanding of climatological processes, familiarity with the data, understanding of statistical methods) and the intention to tell the truth (and thus, for instance, to carefully consider alternative hypotheses).

With this in mind, I posit following thesis:

> **Default Trust (DT).** If scientist A with competence in field *F* makes an assertion φ, then scientist B is believes φ, unless B has an honest disagreement φ due to an incompatible prior belief φ'.

The trust is 'default' in the sense that the trustworthiness of A is not called into question: if A possesses the right type of competence, A can be trusted because A's intention to tell the truth is not doubted. Default trust does not imply agreement: B can withhold high credence in φ if φ is incompatible with a prior belief φ' of B.

DT can be read both normatively and descriptively. A normative reading of it accepts that in reality the expectations scientists have of each other may not be accurately described as 'default trust'. For instance, a sensational but questionable assertion φ that enhances A's career could be distrusted by colleagues. Yet, in a normative reading of DT, default trust between scientific colleagues would be desirable for the scientific ethos.

This paper, unless otherwise specified, will primarily be concerned with the descriptive reading: DT between scientific colleagues characterizes a number of core scientific practices. This descriptive reading does not imply that the scientific ethos is *only* defined by DT – there is room for lack of trust and distrust, under certain circumstances. It holds that the activity of scientific research is characterized by considerable DT.

Take for instance peer-review. As is often acknowledged, the peer-review system is not designed to detect intentional fraud (Crocker and Cooper 2011; Horbach and Halffman 2018). When peer-reviewer B evaluates an assertion φ by author A, B does not necessarily have a way of detecting falsification or fabrication by analysis of the manuscript alone. Image manipulation (of e.g. Western blots) or statistically unlikely patterns in the raw data can be detected; nonetheless, high-profile cases of repeated fabrication went undetected by peer review (such as Diederik Stapel, cf. discussion in Crocker and Cooper 2011). The peer-reviewer can only primarily check the soundness of the manuscript: possible errors in the methodology or reasoning, or check the various assertions against the background of his or her own beliefs. In other words, primarily honest errors will be checked for.

Collaborations are also impossible without default trust. Consider a collaboration between two scientists, A and B, of different specializations (or competences), where B uses A's analysis and conclusion φ to support further analysis. Then B must ultimately trust A's assertion φ unless B wants to redo A's work. Depending on the degree to which φ was unexpected for B, B may of course check in with A for honest errors, and whether various of A's implicit sub-assertions $\varphi_1, \varphi_2, \ldots$ actually imply φ, but at some point B must trust A's work and will not be able to check everything without actually redoing A's work. In this sense, trust is necessary to collaboration: without trust, collaboration becomes literally impossible, since one partner must do or redo all the work.

These core practices illustrate how communication between scientists is permeated by default trust, and that, were such default trust not justified in most cases, then many core scientific practices including peer-review and collaboration would need to be abandoned. Since it is not obvious how a competitive, credit-maximizing model of scientific endeavor can explain such justified default trust, the justified default trust among scientists can be interpreted as an expanandum that should be accounted for.

For the following we will be focusing attention on the question (in connection to credit-maximization): when is default trust justified? Here it is important to distinguish between two ways in which B's trust in scientist A's assertion can be undermined. The first is by reasons to believe that A made an (honest) error in the experimental design, data collection, or data analysis. Thus the assertion φ may not be compatible with the other agent's existing (high credence) beliefs, prompting skepticism towards φ. Note that such reasons can undermine trust in whole fields F, for instance by if it becomes known that a whole field is suffering from widespread methodological problems (see e.g. Sorkin et al. 2016). However, such undermining reasons do not undermine default trust: A's intention to tell the truth is not doubted.

The second way, and more relevant for purposes here, is how trust can be undermined is by learning about the intentions of the scientist for asserting φ. For instance, if scientist A claims that 'smoking does not cause lung cancer' and scientist B finds out that the scientist is being funded by a tobacco company, this not only undermines any trust B might have had in φ, but also undermines B's default trust in A.

3. Credit-Maximizing Norms of Assertion

What is particularly pernicious or disturbing about the problems of sloppy science – the widespread cutting of corners – is that it suggests a widespread culture of scientists putting career over the truth, and hence presents a ubiquitous defeater for the default trust in any scientist. In fact, scientists long have reported that trust is undermined by an incentive structure that actively promotes competition and that is largely based on metrics (Anderson et al. 2007). The question I will consider is: can the credit-maximizing incentive structure be reformed in such a way that default trust is safeguarded?

I will approach this question in the following way: how can the norm of assertion of a credit-maximizing scientist be manipulated by incentivizing replication, such that default trust in that scientist's assertions is justified? The norm of assertion can be stated as follows:

> **Credit-Maximizing Norm of Assertion**. Scientist A will choose to assert φ out of an associated set of possible assertions Φ when φ maximizes the expected credit function C.

Here φ's associated set of possible assertions is defined as $\Phi = \{\varphi, \varphi', \varphi'', \ldots\}$, where the various $\varphi^{(i)}$ are variations, sometimes minute, of the same basic idea, but with different, sometimes radically different, expected pay-off or credit $C(\varphi^{(i)})$.

Note that this norm of assertion is very unlike the norms of assertion traditionally defended by epistemologists (Williamson 2002), which for instance state that an agent can only assert φ when the agent knows φ, or has a high credence in φ. Within a credit maximizing model, it may be 'rational' for a scientist A to assert φ even though A does not know φ, and may even have a low credence in φ. The question of how this norm of assertion precisely overlaps or contrasts with other norms of assertion defended by epistemologists (e.g. Lackey 2007) is outside the scope of this paper.

Stating that such a norm of assertion is rational (relative to a credit-maximizing framework) does not mean that it is necessarily desirable. An unchecked growth in assertions with low credence would mean the death of science, since discourse would be flooded with low credence (and likely false) statements. Hence in a credit maximization model, there needs to be a correction mechanism that disincentivizes low-credence assertions, and the main mechanism that is considered today is replication research.

To further operationalize this norm of assertion, I will expand on Heesen's model of how credit-maximizing scientists should balance speed of output with replicability (Heesen 2018). Assertions that cannot be replicated have negative expected credit (e.g., through reputation loss); yet there is a tradeoff between the credence in a publication (and its replicability) and the speed of publication. Hence Heesen uses following expected credit function:

$$C(p) = c_a \beta p \lambda(p) + c_e \alpha (1-p) \lambda(p)$$

p = scientist's credence that publication is accurate (and also the credence that the article is replicable)
$\lambda(p)$ = expected speed for a publication of replicability p
α = probability of acceptance of erroneous article
β = probability of acceptance of accurate article
c_e/c_a = average credit accrual with erroneous/accurate article

Note that Heesen's model, the 'replicability'[2] of an assertion is scientist A's credence in an assertion, and not the actual replicability: the assumption is thus that this subjective reproducibility closely adheres to actual reproducibility (for if not, and if the scientists' own estimates of replicability were no indication of actual replicability, credit maximization would not be a very good strategy to actually maximize credit).

The credit-maximizing norm of assertion based on Heesen's expected credit function can be described as follows. A scientist must decide between a set of possible assertions $\Phi = \{\varphi, \varphi', \varphi'', \ldots, \varphi^{(n)}\}$, where φ has the lowest replicability (but requires the least supporting work to assert) and $\varphi^{(n)}$ has the highest replicability (and requires the most supporting work to assert). The norm of assertion is *not* to choose to assert the $\varphi^{(i)}$ in which the scientist has highest credence, but rather, to assert the $\varphi^{(j)}$ that has the best trade-off between replicability and speed of publication (and thus highest expected credit).

[2] Heesen uses the term 'reproducibility', but given sentences like "The reproducibility of scientific research is a cornerstone of the scientific method. If science is to discover general laws or principles, it should not matter who tests them, or when, or where.", he seems to be referring to replicability according to NAS's definition (see n1).

This credit-maximizing norm does not necessarily undermine default trust. In fact, it could be reinterpreted as describing a form of practical judgment: attempting to do the best research one can, without succumbing to the temptations of perfectionism. Perfectionism in research describes how a researcher ekes out marginal improvements in accuracy at great cost, thus sabotaging future research. From an ethical perspective, there is nothing necessarily un-integrous about avoiding perfectionism. It is still about doing the best research one can, but considered over a longer time-scale instead of one publication at a time.

However, scientists' credit-maximizing incentive structure is more complex than a trade-off between accuracy and speed, and Heesen's model would need to be expanded to map the issues concerning default trust. Heesen assumes that every assertion will be subject to replication research. However, in reality there are too many assertions to be checked: by some estimates (Ware and Mabe 2015), 2.5 million articles were published in 2014, with a historical growth rate of 3%, so by this logic about 3 million articles will be published in 2020. Not all original assertions can be subjected to replication research. The normative guidelines on replication also reflect this reality: a recent normative guideline on replication research explicitly recommends replication researchers prioritizing those assertions when the results from replication will have an "major impact on scientific knowledge" (KNAW 2018). So to put it more crudely: do not attempt to replicate insignificant assertions. Hence I posit the following additional factor influencing the credit-maximizing norm:

> **Significance of an assertion φ.** Novel, important, or surprising assertions gain more attention than trivial or wholly expected assertions, and are more likely to be the target of replication research.

A further complication I would like to introduce is the fact that replication studies do not always give clear answers (Gilbert et al. 2016). One can submit an assertion φ to replication research, and subsequently not be able to decide whether φ has been confirmed or falsified. This is especially the case where direct replications (where all necessary elements of a procedure are replicated, but with different data) are not possible, leaving only conceptual replications (where the procedure is varied). While it a complex and ongoing question how replications should be conducted (for extensive discussion, see Zwaan et al. 2018), it is safe to say that some assertions are more falsifiable by replication studies than others, and that this falsifiability is relatively independent of the significance of the assertion. Hence I posit a third factor influencing the credit-maximizing norm:

Falsifiability of an assertion φ. Some assertions can be easily confirmed or falsified by replication studies, whereas for other assertions, especially those relying on complex data, replication studies do not either confirm or falsify the original assertion.

In sum, the set of possible assertions Φ can be mapped out on a three-dimensional space where the axes are: **significance, falsifiability,** and **accuracy**. These three dimensions determine either whether an assertion will be subjected to replication research at all (significance), and the probability the replication research will yield a clear confirmation, or a clear falsification, or neither (falsifiability and accuracy).

This means that four scenarios must be distinguished with regards to the fate of an assertion with regards to replicability. (1) The assertion is conclusively successfully replicated, with probability $P(\uparrow)$ and credit accrual $C_{\uparrow}$. (2) The assertion conclusively fails to replicate, with probability $P(\downarrow)$ and credit accrual of $C_{\downarrow}$. (3) The assertion does not either conclusively replicate, nor is conclusively falsified (not replicated), with probability $P(\leftrightarrow)$ and credit accrual $C_{\leftrightarrow}$. (4) The assertion is not subjected to a replication study, with probability $P(0)$ and credit accrual C_0.

As does Heesen (2018), I will assume that $C_{\uparrow}$ is the largest value, and $C_{\downarrow}$ the smallest. In addition, I will assume that $C_{\leftrightarrow}$ is larger than C_0, because an assertion that is not deemed significant to replicate will typically only be published in lower-ranking journals, whereas some results that cannot be conclusively reproduced find their way into a high-ranking journal (as documented by Brembs 2018).

With these four scenarios, the expected credit function of an assertion φ made by scientist A becomes :

$$C_A(\varphi) = C_{\uparrow} P(\uparrow) + C_{\downarrow} P(\downarrow) + C_{\leftrightarrow} P(\leftrightarrow) + C_0 P(0) \qquad (*)$$

Here the exogeneous structural incentive for replication research (e.g., funding, badges, or awards) is inversely correlated with the probability an assertion of average significance is not subjected to a replication study. (If an assertion is very significant (or insignificant), the probability of being subjected to replication may be 1 (or 0) regardless of the strength of the structural incentive).

This function depends on three independent variables[3], and thus the topology of extrema of C_A can be considerably more complex than that in Heesen's model (which had one independent variable, namely the replicability of an assertion). In other words, there is no single way to maximize credit. Scientist A can choose between two possible assertions that are more

[3] The four probability values are constrained by $P(\uparrow) + P(\leftrightarrow) + P(\downarrow) + P(0) = 1$

and less likely to be successfully replicated, but can also choose between assertions that are more and less falsifiable and more and less significant.

In a culture of credit-maximization, where a norm of assertion based on credit function C_* is common knowledge, A asserts φ when $C_A(\varphi)$ is maximal, and B knows that A only asserts φ when $C_A(\varphi)$ is maximal. C_A and B's credit function C_B may not necessarily be maximized by the same assertion φ, for instance in some collaborations where B is lead author and is taking most responsibility. Hence the question: can, in a culture where it is common knowledge credit-maximization, incentivizing replication research safeguard B's default trust in A?

Despite C_* not having obvious maxima, what is perhaps more realistic of actual reasoning processes is how credit-maximizing solutions can be attained iteratively instead of analytically, by a search strategy consisting of a series of decision. This search thus consists of decision tree an agent will follow in a quest to maximize credit. Credit maximization, and not truth-telling per se, determines the calculus behind every decision.

As an example of a rather simple search strategy, consider the following: the scientist first starts with the most significant assertion, which maximizes the largest credit accrual type $C_\uparrow$, and from there goes down the ladder of pay-offs guided by $C_\uparrow > C_\leftrightarrow > C_0 > C_\downarrow$. In more detail:

(a) Look for a maximally significant φ (minimizing $P(0)$).

(b) If, with some effort, it can be made sufficiently reproducible (maximizing $P(\uparrow)$ while minimizing $P(\leftrightarrow)$ and $P(\downarrow)$) then assert. This is the ideal, maximum pay-off scenario.

(c) If, with another additional effort, φ unfalsifiability can be maximized, then assert.

(d) If not consider a next assertion φ' in Φ with slightly lower significance, and either (when potential pay-off is high enough) go through the same process again, or else consider the project to be a failure and move on to the next.

(e) If stopping without assertion is not an option (e.g. due to large investment in starting up the project), then in the worst case scenario the search process is stopped when arriving at a maximally significant assertion $\varphi^{(n)}$ that just about insignificant enough that it will not attract any attention. Such a norm can be asserted.

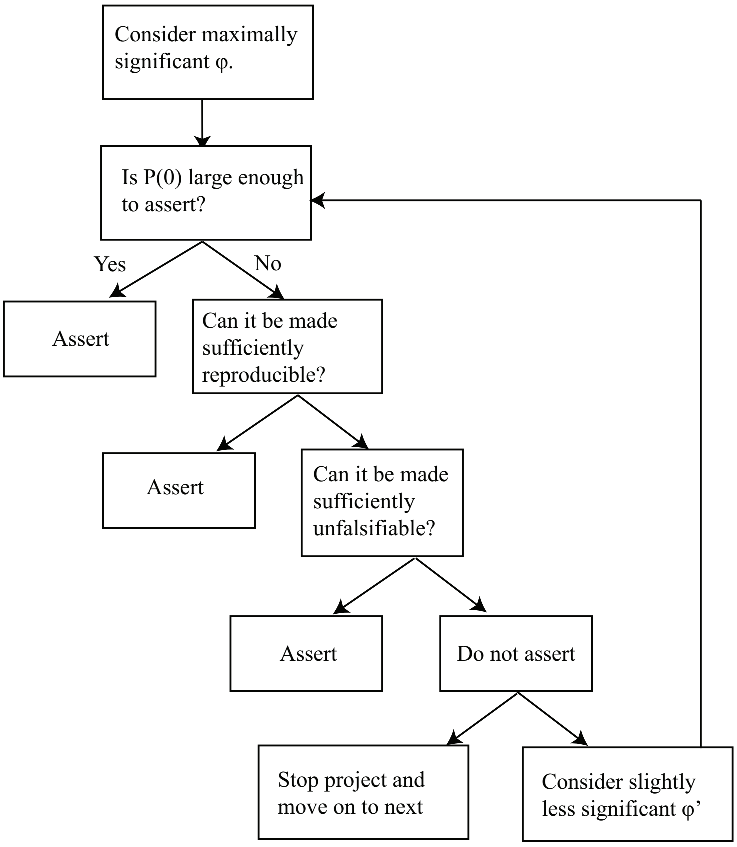

Figure 1: An example of a decision tree a credit-maximizing scientist could use to search Φ for an assertion with maximal expected credit.

Note that this decision-making process makes two, reasonably plausible, assumptions. The first is that a scientist must also decide whether it is worth continuing the search, or in other words, whether the extra additional investment needed for continuing the search is smaller than the expected payoff. The second is that peer review does not present an obstacle for the assertion of φ, so that if φ is unfalsifiable, that the scientist is sufficiently experienced to hide the unfalsifiability. If φ is not highly significant, then the assertion will be published in a lower-ranked journal.

4. Credit Maximization and Trust

The credit-maximizing reasoning sketched in Figure 1 is put forward primarily for illustrative & heuristic purposes. It is linear and highly simplified, and does not necessarily reflect, for instance, how researchers may simultaneously consider multiple possible assertions. What is important is that in a credit-maximizing culture, scientist B knows that A followed *some* credit-maximizing decision-making tree, and thus A could assert φ without knowing φ or even having a high credence in φ.

When is scientist B justified in defaultly trusting an assertion φ of scientist A? Recall that default trust can be undermined by beliefs about the intentions of A. B has least reason to not to trust A when B receives indications that $P(\uparrow)$ is maximized. Since $P(\uparrow)$ is also A's credence in φ, this is cognitively inaccessible to B; thus B must estimate $P(\uparrow)$ by other means. And when replication research is incentivized this is possible. The significance of φ means that φ is likely to be submitted to replication research ($P(0)$ is low). The falsifiability of φ means that a replication study of φ will be conclusive ($P(\leftrightarrow)$ is low). Hence, perceptions of the significance and falsifiability of A's assertion φ would be good grounds to believe that A's asserts φ while also having a high credence in φ. Given these considerations, credit maximization gives support for the following sense of qualified trust:

> **Justified Qualified Trust.** If a scientist A specialized in field *F* makes an assertion Φ, then scientist B is only justified in believing φ when φ is highly significant and clearly falsifiable.

How can significance and falsifiability be estimated by scientist B? The former is relatively straightforward: since *B* is an expert in the field, *B* can directly infer, from his or her background knowledge, whether an assertion φ is significant or not. Hence B can assume that *if* A is making a highly significant assertion, A will know that it will attract replication research, and will want to minimize the probability of falsification.

Falsifiability is more difficult to assess. Lack of falsifiability can sometimes be hidden by the author: one can always fail to mention some crucial limiting assumption that will prevent any conclusive confirmation or falsification (see again Zwaan et al. 2018). However, falsifiability of an assertion can be explicitly signaled by an author. For instance, the author can be minute in describing procedural detail, thus giving explicit instructions how to replicate the findings. Or, the author can share the raw data on which the assertion was based. These, not coincidentally, are some of the core measures proposed to increase reproducibility of research (Munafò et al. 2017). Such measures, perhaps more importantly, also increase trust in a credit-maximizing world since they signal falsifiability.

5. Conclusion

Default trust is necessary for a number of core scientific practices, such as peer review or collaboration. In a world where assertions can not only be true or not, but also significant or not and easily falsifiable or not, it becomes impossible to assume that scientists can be sufficiently incentivized to only make assertions that can be maximally replicated. In a complex credit function, there are many different strategies to maximize credit, and not all of those strategies align with truth-telling. This means that default trust, in a credit-maximizing culture, cannot be justified.

I would also like to caution against the normative inference that any absence of default trust is necessarily a bad thing. It may be ultimately beneficial for the scientific enterprise if the scientific community has some elements of the jungle, where the genuine and the fake need to be sifted, or where scientists jostle to make assertions of significances, even though this can occasionally lead to untrustworthy science.

Nonetheless, what the argument does suggest is that in an extreme culture of credit-maximization, default trust even within collaborations would no longer be justified: scientist B would need to be calculating whether it is in the best credit-maximizing interests of scientist A to do good work and make trustworthy assertions. Scientist B would need to be vigilant of the situation where scientist A could get away with sloppy work. This would no longer be a relationship of trust, and it is clear that tight-knit and efficient collaborations would no longer be possible.

References

Anderson, Melissa S., Emily A. Ronning, Raymond De Vries, and Brian C. Martinson. 2007. “The Perverse Effects of Competition on Scientists’ Work and Relationships.” *Science and Engineering Ethics* 13 (4): 437–61. https://doi.org/10.1007/s11948-007-9042-5.

Baker, Monya. 2016. “Is There a Reproducibility Crisis?” *Nature News* 533 (7604): 452–54. https://doi.org/10.1038/533452a.

Brembs, Björn. 2018. “Prestigious Science Journals Struggle to Reach Even Average Reliability.” *Frontiers in Human Neuroscience* 12 (February): 37. https://doi.org/10.3389/fnhum.2018.00037.

Collier, Roger. 2015. “Scientific Misconduct or Criminal Offence?” *CMAJ : Canadian Medical Association Journal* 187 (17): 1273–74. https://doi.org/10.1503/cmaj.109-5171.

Crocker, Jennifer, and M. Lynne Cooper. 2011. “Addressing Scientific Fraud.” *Science* 334 (6060): 1182–1182. https://doi.org/10.1126/science.1216775.

Fanelli, Daniele. 2018. “Opinion: Is Science Really Facing a Reproducibility Crisis, and Do We Need It To?” *Proceedings of the National Academy of Sciences* 115 (11): 2628–31. https://doi.org/10.1073/pnas.1708272114.

Gilbert, D. T., G. King, S. Pettigrew, and T. D. Wilson. 2016. “Comment on ‘Estimating the Reproducibility of Psychological Science.’” *Science* 351 (6277): 1037–1037. https://doi.org/10.1126/science.aad7243.

Gorgolewski, Krzysztof J., Thomas Nichols, David N. Kennedy, Jean-Baptiste Poline, and Russell A. Poldrack. 2018. “Making Replication Prestigious.” *Behavioral and Brain Sciences* 41. https://doi.org/10.1017/S0140525X18000663.

Hawley, Katherine. 2012. *Trust: A Very Short Introduction*. First edition. Very Short Introductions 325. Oxford: Oxford University Press. https://books.google.be/books/about/Trust_A_Very_Short_Introduction.html?id=lZIuLVq8-JEC&redir_esc=y.

Heesen, Remco. 2018. "Why the Reward Structure of Science Makes Reproducibility Problems Inevitable." *The Journal of Philosophy* 115 (12): 661–74. https://doi.org/10.5840/jphil20181151239.

Horbach, S. P. J. M., and W. (Willem) Halffman. 2018. "The Changing Forms and Expectations of Peer Review." *Research Integrity and Peer Review* 3 (1): 8. https://doi.org/10.1186/s41073-018-0051-5.

Kitcher, Philip. 1990. "The Division of Cognitive Labor." *The Journal of Philosophy* 87 (1): 5. https://doi.org/10.2307/2026796.

KNAW. 2018. *Improving Reproducibility in the Empirical Sciences*. Amsterdam: KNAW.

Lackey, Jennifer. 2007. "Norms of Assertion." *Noûs* 41 (4): 594–626. https://doi.org/10.1111/j.1468-0068.2007.00664.x.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1 (1): 0021. https://doi.org/10.1038/s41562-016-0021.

NAS, (National Academies of Sciences, Engineering, and Medicine). 2019. *Reproducibility and Replicability in Science*. Washington, D.C.: National Academies Press. https://doi.org/10.17226/25303.

Sorkin, Barbara C, Adam J Kuszak, John S Williamson, D Craig Hopp, and Joseph M Betz. 2016. "The Challenge of Reproducibility and Accuracy in Nutrition Research: Resources and Pitfalls." *Advances in Nutrition* 7 (2): 383–89. https://doi.org/10.3945/an.115.010595.

Ware, Mark, and Michael Mabe. 2015. "The STM Report: An Overview of Scientific and Scholarly Journal Publishing," 181.

Williamson, Timothy. 2002. *Knowledge and Its Limits*. Oxford University Press.

Zwaan, Rolf A., Alexander Etz, Richard E. Lucas, and M. Brent Donnellan. 2018. "Making Replication Mainstream." *Behavioral and Brain Sciences* 41. https://doi.org/10.1017/S0140525X17001972.