Incentivizing Replication is Insufficient to Safeguard Default Trust

Hugh Desmond

Forthcoming in Philosophy of Science

Abstract

Philosophers of science and meta-scientists alike 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; Smaldino and McElreath 2016; Heesen 2018; O'Connor 2019).

Yet in applied ethics, sociology of science, and to some extent also science policy, an 'ethical approach' informs research on social structures of science: individual scientists are predominantly understood to be agents capable of acting contrary to incentive structures, and concerned with the virtues and ideals of research integrity, such as honesty, respect, or reliability (Carvalho 2017; Desmond 2020; ALLEA 2017; Forsberg et al. 2018; Pennock 2019).

Philosophers of science and meta-scientists do not openly dismiss the value of research integrity as such. In economic approaches, the capacity for ethical decision-making is implicitly assigned to policy-makers (or the "philosopher-monarchs" in Kitcher 1990): these agents take the long view and decide on how incentives should be designed. However, the question arises whether the ethical approach to the decision-making of individual scientists brings anything to the explanatory table that cannot be covered by the economic approach. For instance, individual concern with honesty could be reduced to 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).

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

The importance of this question extends to economic approaches to science policy. 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 focus on *default trust between scientists* as something that cannot be engineered by incentive reform. 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 – peer review and collaboration are considered 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 that such problems are widespread (Fanelli 2018) – 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. The question then becomes: can incentivizing replication safeguard default trust? Using an expanded version of Heesen's model of when replicable (or trustworthy) assertions maximize credit (section 3), I show that, no matter how much replication research is incentivized, default trust cannot be justified in a culture of credit-maximization (section 4).

One upshot is that the economic model cannot account for an important explanandum concerning scientific practice (i.e., default trust). Another, more normative upshot is that a culture of extreme credit-maximization ultimately erodes default trust between scientists, and even gives rise to a 'default lack of trust'. This shows why it would be misguided for policy efforts to focus on incentive reform alone.

---

[1] I will adopt the National Academy of Science's definitions of reproducibility (roughly: obtaining consistent results by redoing the same analysis on the same data) and replicability (roughly: 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 a summary, see Hawley 2012), trust in an assertion φ is typically understood to depend on the asserting agent's intention to tell the truth and competence to know that φ. If the agent either lacks knowledge or is dishonest, trust would be misplaced. 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 believes φ, unless B has an honest disagreement with 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, then A can be trusted because A's intention to tell the truth is not doubted. Note that 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. In a normative reading, DT is an ideal for the scientific ethos, even though actual behavior may fall short. For instance, a sensational but questionable assertion φ that enhances A's career could be distrusted by colleagues. Thus in reality, the stances scientists have towards each other may not necessarily correspond to 'default trust'.

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 for distrust, under certain circumstances. It just holds that the activity of scientific research is characterized by considerable DT, and moreover, core scientific practices would not be possible without 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; Serge P. J. M. 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 have gone 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 whether the manuscript's assertions cohere with the reviewer's background beliefs. In other words, the reviewer primarily checks for honest errors.

Collaboration is another core scientific practice that would be impossible without default trust. Consider a collaboration between two scientists A and B, where A and B have different specializations (i.e., competences) and where B uses A's analysis and conclusion φ to support further analysis. Then B must ultimately trust A's assertion φ unless B would want 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. Nonetheless, at some point B must trust A's work and will not be able to check everything – unless, B starts redoing A's work, but then one no longer speak of a "collaboration" between A and B. In this sense, default trust is a necessary precondition for collaboration.

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 posited 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 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, 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 A 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 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 metric-based incentive structures that actively promote competition (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? Or is this a false expectation?

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 must 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 $\Phi = \{\varphi, \varphi', \varphi'', \dots\}$, 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 (e.g. Williamson 2000), 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.

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. The main mechanism that is considered today is replication research.

To further operationalize this norm of assertion, I will expand on Remco Heesen's model of how credit-maximizing scientists should balance speed of output with replicability (Heesen 2018). In this model, articles or assertions that cannot be replicated have lower expected credit (e.g., through reputation loss), but high-credence (and likely replicable) assertions require more investment and thus are made at a lower frequency. In particular, 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 assertion is accurate (and also the credence that the assertion is replicable)
$\lambda(p)$ = expected speed for constructing an article 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) publication

Note that in 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 expected replicability closely adheres to actual replicability. (This is a reasonable assumption, because, if it were false, and if the scientists' own estimate of replicability was not in any way a reliable indicator of actual replicability, then aiming at credit maximization would not be a good strategy to 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 research work in order to assert) and $\varphi^{(n)}$ has the highest replicability (and requires the most research work in order 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).

This credit-maximizing norm does not necessarily undermine default trust. In fact, it could be reinterpreted as a form of practical wisdom: attempting to do the best research one can without succumbing to perfectionism. When a researcher gives in to perfectionism, in marginal improvements in accuracy are eked out at great cost, thus sabotaging future research and overall credit. From an ethical perspective, there is nothing necessarily non-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.

To map the issues concerning default trust, we need a more complex and realistic model of credit maximization. First of all, not all original assertions are subjected to replication research: by and large, only significant assertions are. This is inevitable: by some estimates

[2] Heesen uses the term "reproducibility", but his use seems to correspond to "replicability" according to NAS's definition (see n1; compare with Heesen 2018, p. 663).

(Ware and Mabe 2015), 2.5 million articles were published in 2014, and with a historical growth rate of 3%, this would imply over 3 million published articles will be published in 2021. Some selection is needed.[3] Normative guidelines on replication also reflect this reality: a recent guideline explicitly recommends replication researchers to prioritize those assertions when the results from replication will have an "major impact on scientific knowledge" (KNAW 2018). Or to put it more crudely: do not bother with replicating insignificant assertions. Hence the credit-maximizing scientist must take the significance of the assertion into consideration:

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.

The second element that should be present in a more realistic model of credit maximization reflects the fact that replication studies do not always give clear answers (Gilbert et al. 2016). Thus, 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 one can posit a second additional 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).

[3] In 2014, the number of researchers worldwide was between 7 and 9 million, of which only 1% published more than one article per year. It seems reasonable to assume that a large number of published articles will never attract the attention of replication researchers.

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

As in Heesen (2018), I will also 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 may still 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:

$$C_A\,(\varphi) = C_{\uparrow}\,P(\uparrow) + C_{\downarrow}P(\downarrow) + C_{\leftrightarrow}P(\leftrightarrow) + C_0P(0) \qquad (*)$$

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).

The function C_A depends on three independent variables[4], and thus the topology of extrema of C_A is likely to be such that there may not be any single way to maximize credit. In other words, there may be different credit-maximizing strategies. Nonetheless, instead of credit-maximization as an analytic calculation (as is possible in Heesen's model), one can also think of credit-maximization as an iterative search process. This search thus consists of decision tree an agent will follow in a quest to maximize credit.

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 $P(0)$).

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

(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.

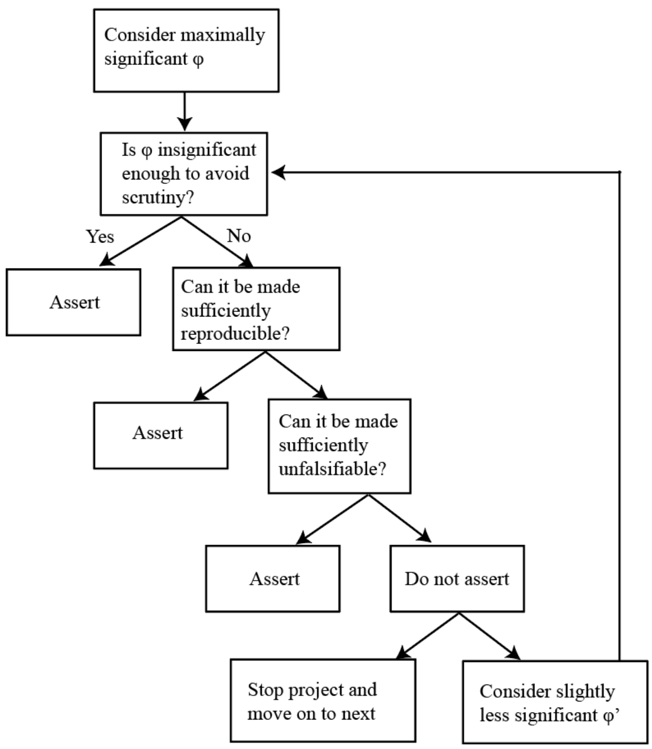

Figure 1: An example of a decision tree a credit-maximizing scientist could use to search *Φ* for an assertion with maximal expected credit.

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 its unfalsifiability. If φ is not highly significant, then the assertion will be published in a lower-ranked journal.

4. Credit Maximization and Trust

The illustrative credit-maximizing reasoning sketched in Figure 1 is linear and highly simplified, and does not reflect, for instance, how researchers may simultaneously consider multiple possible assertions. The main lesson to be drawn is that, in a research culture characterized by credit-maximization, another scientist B can know that A has followed *some* credit-maximizing decision-making tree prior to assertion. In other words, the function C_A is *common knowledge* and that the associated norm of assertion is a social norm (cf. Bicchieri 2016): A asserts φ only when A believes C_A (φ) is maximal; B knows that A only asserts φ when C_A (φ) is believed to be maximal; A knows that B knows; and so on. Given such a credit-maximizing culture, B may not know which precise search strategy A employs, but B can know that A is swayed by credence (or accuracy), significance, and falsifiability and hence that A could assert φ without knowing φ or even having a high credence in φ.

This may sound like a defeater for trust, but that is why replication research is so important: it makes it difficult for A to get away with low-credence assertions. Hence, if replication research is incentivized (lowering $P(0)$), B knows that A is less likely to assert φ without A's credence in φ, $P(\uparrow)$, being high. Thus incentivizing replication research can help maintain B's trust in A's assertions.

So while incentivizing replication may increase trust, the question becomes whether it is sufficient to maintain (justified) *default trust* in a culture defined by a norm of credit-maximization as described by C_A. Recall that default trust is a type of trust where the truth-telling intentions of A are not questioned: A's assertion φ is evidence of A's high credence in φ. However, default trust can be undermined by beliefs about the intentions of A. For instance, B knows that, if A's assertion φ is relatively insignificant, then φ is unlikely to be submitted to replication research ($P(0)$ is high). Similarly, B knows that, if φ seems difficult to falsify, then a replication study of φ may not be conclusive ($P(\leftrightarrow)$ is high). In both these cases, B knows that A can conceivably assert without high credence. In other words, B can only safely infer A's high credence when A makes a significant and falsifiable assertion. Given these considerations, a credit-maximizing norm of assertion only 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 often 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 by B. Failure to replicate surprisingly often does not imply that the assertion φ is false (see again Zwaan et al. 2018). Nonetheless, the 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): perhaps more importantly, such measures also increase qualified trust in a credit-maximizing world.

5. Conclusion

In a world where assertions can not only be true or false, but also significant or insignificant, or easily falsifiable or not, it becomes impossible to assume that scientists can be sufficiently incentivized to only make assertions that can be maximally replicated. Not all strategies for success align with truth-telling. This means that default trust, in a sufficiently sophisticated credit-maximizing culture, cannot be justified.

There is nothing inherently wrong about the scientific enterprise having some elements of the jungle, with scientists jostling to make assertions of significances, even though this can occasionally lead to untrustworthy science. Nonetheless, given how crucial default trust is for basic scientific practices like peer-review or collaboration, it would seem important to place limits on what one can expect from efforts to engineer incentive structures.

Acknowledgments
The author would like to acknowledge the audience of an online workshop on the replicability crisis organized by Olivier Leclerc and Stéphanie Ruphy, as well as Stijn Conix, Remco Heesen, and Liam Kofi Bright for their helpful comments.

References

ALLEA, All European Academies. 2017. *The European Code of Conduct for Research Integrity*. http://www.allea.org/wp-content/uploads/2017/04/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf.

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.

Bicchieri, Cristina. 2016. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press.

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.

Carvalho, Teresa. 2017. "The Study of the Academic Profession – Contributions from and to the Sociology of Professions." *Theory and Method in Higher Education Research* 3: 59–76. https://doi.org/10.1108/S2056-375220170000003004.

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.

Desmond, Hugh. 2020. "Professionalism in Science: Competence, Autonomy, and Service." *Science and Engineering Ethics* 26 (3): 1287–1313. https://doi.org/10.1007/s11948-019-00143-x.

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.

Forsberg, Ellen-Marie, Frank O. Anthun, Sharon Bailey, Giles Birchley, Henriette Bout, Carlo Casonato, Gloria González Fuster, et al. 2018. "Working with Research Integrity—Guidance for Research Performing Organisations: The Bonn PRINTEGER Statement." *Science and Engineering Ethics* 24 (4): 1023–34. https://doi.org/10.1007/s11948-018-0034-4.

Gilbert, Daniel T, Gary King, Stephen Pettigrew, and Timothy 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*. 325. Oxford: Oxford University Press.

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, Serge P. J. M., and 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.
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.
O’Connor, Cailin. 2019. “The Natural Selection of Conservative Science.” *Studies in History and Philosophy of Science Part A* 76 (August): 24–29. https://doi.org/10.1016/j.shpsa.2018.09.007.
Pennock, Robert T. 2019. *An Instinct for Truth: Curiosity and the Moral Character of Science*. Cambridge, MA: The MIT Press.
Smaldino, Paul E., and Richard McElreath. 2016. “The Natural Selection of Bad Science.” *Royal Society Open Science* 3 (9): 160384. https://doi.org/10.1098/rsos.160384.
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.
Strevens, Michael. 2006. “The Role of the Matthew Effect in Science.” *Studies in History and Philosophy of Science Part A* 37 (2): 159–70. https://doi.org/10.1016/j.shpsa.2005.07.009.
Ware, Mark, and Michael Mabe. 2015. “The STM Report: An Overview of Scientific and Scholarly Journal Publishing.” http://digitalcommons.unl.edu/scholcom/9.
Williamson, Timothy. 2000. *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.