Why the Reward Structure of Science Makes Reproducibility Problems Inevitable*

Abstract

Recent philosophical work has praised the reward structure of science, while recent empirical work has shown that many scientific results may not be reproducible. I argue that the reward structure of science incentivizes scientists to focus on speed and impact at the expense of the reproducibility of their work, thus contributing to the so-called reproducibility crisis. I use a rational choice model to identify a set of sufficient conditions for this problem to arise, and I argue that these conditions plausibly apply to a wide range of research situations. Currently proposed solutions will not fully address this problem. Philosophical commentators should temper their optimism about the reward structure of science.

*This is the peer-reviewed and copy-edited (but not journal-formatted) version of the following article: Remco Heesen, "Why the Reward Structure of Science Makes Reproducibility Problems Inevitable," THE JOURNAL OF PHILOSOPHY, CXV, 12 (December 2018): 661–74, which has been published in final form at doi.org/10.5840/jphil20181151239. This article may be used for non-commercial purposes only. To contact the author write to remco.heesen@uwa.edu.au. Thanks to Kevin Zollman, Michael Strevens, Stephan Hartmann, Teddy Seidenfeld, Jan Sprenger, Liam Bright, Cailin O'Connor, Seamus Bradley, Conor Mayo-Wilson, Rory Švarc, an anonymous reviewer for this JOURNAL, and audiences at Tilburg University, the National University of Singapore, the Congress of Logic, Methodology and Philosophy of Science in Helsinki, and the Formal Epistemology Workshop in Groningen for valuable comments and discussion. This work was partially supported by the National Science Foundation under grant SES 1254291 and by an Early Career Fellowship from the Leverhulme Trust and the Isaac Newton Trust.

The reward structure of science has been of increasing interest to philosophers. The literature on this subject has focused on the good news: ways in which rewards can contribute to scientific progress.[1] The present contribution nuances this message by highlighting some bad news, in particular that the reward structure gives scientists an incentive to rush into print, which plausibly contributes to reproducibility problems.

A central aim of the philosophical literature on the reward structure seems to be to argue against the view that scientific progress is best served when individual scientists are epistemically rational. A paradigm case is the argument by Kitcher and Strevens that reward-seeking scientists will choose research programs or methodologies in a way that makes for a socially bene-

[1] A number of these papers have appeared in this JOURNAL: Philip Kitcher, "The Division of Cognitive Labor," this JOURNAL, LXXXVII, 1 (January 1990): 5–22; Michael Strevens, "The Role of the Priority Rule in Science," this JOURNAL, C, 2 (February 2003): 55–79; Kevin J. S. Zollman, "The Credit Economy and the Economic Rationality of Science," this JOURNAL, CXV, 1 (January 2018): 5–33. Other optimistic appraisals of the reward structure by philosophers and economists include Michael Polanyi, "The Republic of Science: Its Political and Economic Theory," *Minerva*, I, 1 (Autumn 1962): 54–73; David L. Hull, *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science* (Chicago: University of Chicago Press, 1988); David L. Hull, "What's Wrong with Invisible-Hand Explanations?," *Philosophy of Science*, LXIV, Proceedings (1997): S117–26; Philip Kitcher, *The Advancement of Science: Science without Legend, Objectivity without Illusions* (Oxford: Oxford University Press, 1993); Partha Dasgupta and Paul A. David, "Toward a New Economics of Science," *Research Policy*, XXIII, 5 (September 1994): 487–521; Thomas C. Leonard, "Reflection on Rules in Science: An Invisible-Hand Perspective," *Journal of Economic Methodology*, IX, 2 (2002): 141–68; Thomas Boyer, "Is a Bird in the Hand Worth Two in the Bush? Or, Whether Scientists Should Publish Intermediate Results," *Synthese*, CXCI, 1 (January 2014): 17–35; Thomas Boyer-Kassem and Cyrille Imbert, "Scientific Collaboration: Do Two Heads Need to Be More than Twice Better than One?," *Philosophy of Science*, LXXXII, 4 (October 2015): 667–88; Peter J. Boettke and Kyle W. O'Donnell, "The Social Responsibility of Economists," in George F. DeMartino and Deirdre N. McCloskey, eds., *The Oxford Handbook of Professional Economic Ethics* (Oxford: Oxford University Press, 2016), pp. 116–36; Michael Strevens, "Scientific Sharing, Communism, and the Social Contract," in Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, eds., *Scientific Collaboration and Collective Knowledge: New Essays* (Oxford: Oxford University Press, 2017), pp. 3–33; and Remco Heesen, "Communism and the Incentive to Share in Science," *Philosophy of Science*, LXXXIV, 4 (October 2017): 698–716.

ficial division of labor.[2] Since writers on epistemic rationality focus on what it is rational to believe rather than what it is rational to do,[3] it is not obvious what, if anything, would be the epistemically rational choice of methodology.[4] Kitcher assumes that epistemically rational scientists choose whichever methodology they think has the greatest chance of success and argues that the distribution of scientists over methodologies this produces can potentially be improved by an appropriate reward structure, while Strevens focuses only on showing that reward-seeking scientists can achieve an optimal distribution.

Kitcher, Strevens, and the other authors listed in footnote 1 all use the apparatus of decision and game theory to investigate how rational scientists might respond to various reward structures. Further, they each praise a particular reward structure as incentivizing individual behavior that is good for scientific progress. In doing so these authors take an optimistic stance on the reward structure. They acknowledge that scientists may be motivated by a desire for personal reward (that is, credit or prestige) but then go on to suggest that, somewhat surprisingly, this leads to better outcomes than a hypothetical scientific enterprise populated by high-minded scientists who are indifferent to credit. The reward structure ends up looking much like Adam Smith's invisible hand, guiding self-interested individual scientists to

[2]Kitcher, "Division of Cognitive Labor," *op. cit.*; and Strevens, "Role of the Priority Rule," *op. cit.* Related work on scientists' choice of research program or methodology with less of a focus on rewards includes Michael Weisberg and Ryan Muldoon, "Epistemic Landscapes and the Division of Cognitive Labor," *Philosophy of Science*, LXXVI, 2 (April 2009): 225–52; Kevin J. S. Zollman, "The Epistemic Benefit of Transient Diversity," *Erkenntnis*, LXXII, 1 (January 2010): 17–35; Conor Mayo-Wilson, Kevin J. S. Zollman, and David Danks, "The Independence Thesis: When Individual and Social Epistemology Diverge," *Philosophy of Science*, LXXVIII, 4 (October 2011): 653–77; and Johanna Thoma, "The Epistemic Division of Labor Revisited," *Philosophy of Science*, LXXXII, 3 (July 2015): 454–72.

[3]See, for example, Thomas Kelly, "Epistemic Rationality as Instrumental Rationality: A Critique," *Philosophy and Phenomenological Research*, LXVI, 3 (May 2003): 612–40; and Richard Pettigrew, *Accuracy and the Laws of Credence* (Oxford: Oxford University Press, 2016).

[4]Zollman, "Credit Economy," *op. cit.*, argues that this is an important omission.

socially beneficial choices.[5]

I agree with much of the broader message in this work. There are interesting normative questions to be asked about science that go beyond the traditional ones about rational belief and evidence, and thinking about the reward structure of science is a fruitful source of such questions and potential answers. My aim in this contribution is not to challenge these virtues but rather to temper some of the optimism mentioned above.

The reward structure of science does not always act like an invisible hand. In some situations there is a systematic misalignment between what rational credit-maximizing scientists would do and what would be best for them to do from a social perspective. I illustrate this by studying the question of how much research a scientist should do before publishing her work. I argue that in many cases there will be an incentive to publish quickly, which plausibly contributes to the reproducibility crisis that has recently received significant attention.

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. Thus it is a necessary condition for the acceptability of a particular scientific result that, if some (hypothetical or actual) scientist competently performs the same experiment, it produces the same result.

Especially in medicine and psychology, there has long been "a general impression that many results that are published are hard to reproduce,"[6] which has recently begun to be empirically tested. Two studies by pharmaceutical companies could reproduce less than a quarter of results in cancer biology.[7]

[5]For explicit comparisons of the reward structure of science to an invisible hand, see in particular Hull, *Science as a Process*, *op. cit.*; Hull, "Invisible-Hand Explanations," *op. cit.*; Leonard, "Rules in Science," *op. cit.*; and Polanyi, "Republic of Science," *op. cit.*

[6]Florian Prinz, Thomas Schlange, and Khusru Asadullah, "Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?," *Nature Reviews Drug Discovery*, x, 9 (2011): 712.

[7]*Ibid.*; and C. Glenn Begley and Lee M. Ellis, "Raise Standards for Preclinical Cancer

A large, more systematic study of prominent results in psychology found that less than 40% could be reproduced, while similar studies of social science experiments and experimental philosophy successfully reproduced about 60% and about 70%, respectively.[8]

Low empirical reproducibility rates do not prove by themselves that there is a problem (they could simply be an indication that there are few true discoveries to be made), but my claim is that there is in fact a problem, and it stems from the reward structure of science.

Claim (Rushing into print). Scientists are incentivized to produce more results at the expense of spending more time on the reproducibility of any given result.

The aim of the rational choice model I present below is to establish conditions for this claim to hold. I argue that three basic ingredients are sufficient: first, the fact that speed and reproducibility trade off against each other; second, the fact that scientists get rewarded for publications; and third, the fact that publications depend on peer review, which has to assess the medium- to long-term impact of papers in the short term, and necessarily does so imperfectly.

My analysis differs from those that identify particular journal practices[9] or

Research," *Nature*, CCCCLXXXIII, 7391 (Mar. 29, 2012): 531–33. A more systematic study is currently underway; see Brian A. Nosek and Timothy M. Errington, "Reproducibility in Cancer Biology: Making Sense of Replications," *eLife*, VI (2017): e23383.

[8]Open Science Collaboration, "Estimating the Reproducibility of Psychological Science," *Science*, CCCXLIX, 6251 (Aug. 28, 2015): aac4716; Colin F. Camerer et al., "Evaluating the Replicability of Social Science Experiments in *Nature* and *Science* between 2010 and 2015," *Nature Human Behaviour*, II, 9 (2018): 637–44; and Florian Cova et al., "Estimating the Reproducibility of Experimental Philosophy," *Review of Philosophy and Psychology* (forthcoming), `http://dx.doi.org/10.1007/s13164-018-0400-9`.

[9]Such as "publication bias," a preference for positive or statistically significant results. See P. J. Easterbrook, R. Gopalan, J. A. Berlin, and D. R. Matthews, "Publication Bias in Clinical Research," *The Lancet*, CCCXXXVII, 8746 (Apr. 3, 1991): 867–72; and Matthias Egger and George Davey Smith, "Bias in Location and Selection of Studies," *BMJ*, CCCXVI, 7124 (1998): 61–66.

statistical practices[10] as the only sources of reproducibility problems. I do not deny that these practices exist and contribute to reproducibility problems, or that it would be a good idea to implement the remedies they suggest (such as publishing null results and requiring pre-analysis plans). However, my model does not incorporate these problematic practices and hence shows that the proposed remedies do not suffice to eliminate reproducibility problems. In this sense my analysis is more general, implying that the problem is harder to solve than might otherwise be thought.

I discuss possible remedies in the final section of this paper. I argue that no "nearby" reward structure fully solves this problem. This is the sense in which I temper the optimism of the philosophical literature on the reward structure: whereas for a number of issues, including the choice of methodology, there are (under certain assumptions) reward structures that incentivize socially optimal choices, I argue that the analogous claim for the tradeoff between speed and reproducibility fails to hold.

I. A TRADEOFF BETWEEN SPEED AND REPRODUCIBILITY

Consider a scientist working on a research study. When should she attempt to publish her work? Because I am interested in what the scientist has a credit incentive to do, I assume that *credit* is her only concern in making this decision. This is a *methodological* assumption to isolate the credit incentive.

Since the scientist aims to maximize the amount of credit she accrues per unit time, she prefers to publish quickly rather than slowly (all else being equal): the concern for credit entails a concern for *speed* (to be defined more formally below). At the same time, publishing faster reduces *reproducibility*. By reproducibility I mean, loosely speaking, the likelihood that the result of the research study is reproduced if someone attempts to do so.

[10] Such as data dredging. See Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn, "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, XXII, 11 (November 2011): 1359–66.

This loose definition of reproducibility has two problems. First, what if no one attempts to reproduce the result? And second, what if multiple scientists attempt to reproduce it, with some succeeding and some failing? Since credit is conferred socially, what really matters is the standing of a result in the eyes of other scientists. So I call a scientific result *accurate* if it holds up in the relevant scientific community in the mid-term: either no one attempts to reproduce it, or subsequent studies are taken on balance to reproduce the result. Conversely, I call a result *erroneous* if it does not hold up in the mid-term, that is, if the community deems the result irreproducible. The *reproducibility* of the result is then the scientist's subjective probability, given the evidence gathered at the time of publication, that the result is accurate. This definition should be interpreted broadly, applying to both experimental and non-experimental contributions (for example, a mathematical theorem is considered reproducible if no one discovers a mistake in it).

In the model, the scientist chooses the desired reproducibility $p \in [0, 1]$ *ex ante*. I assume this to be fixed for the duration of the study. That is, the scientist works on her study until she thinks her result has at least probability p of holding up in the community, at which time she publishes.

Reproducibility takes time. This is reflected in the model by the *speed function* λ. The value $\lambda(p)$ represents the scientist's expected speed if the desired reproducibility is p, that is, the number of studies "like this one" that the scientist would expect to complete per unit time (see Figure 1). So $\mu(p) = 1/\lambda(p)$ is the (*ex ante*) expected time until completion of the study.

Reducing reproducibility (lowering p) allows the scientist to publish faster. "Rushing" the work in this way could mean that the scientist ends the study sooner (gathering less evidence), or it could mean that the scientist tries to gather the same amount of evidence more quickly (potentially making mistakes). The present model is not intended to investigate incentives related

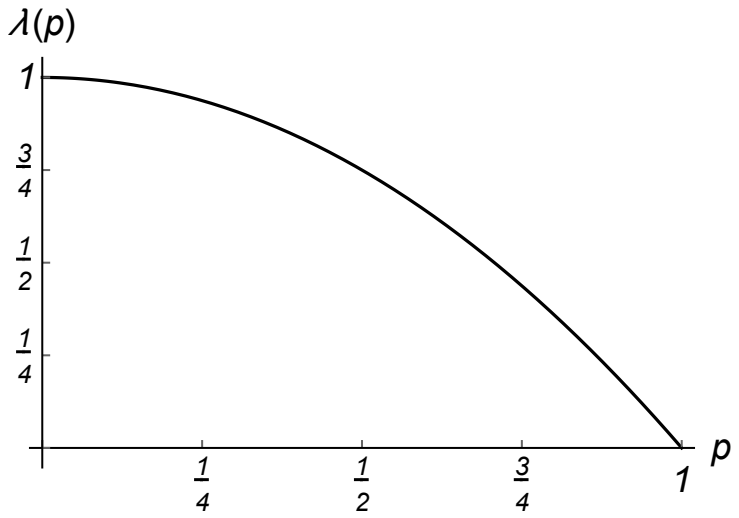

Figure 1: p and λ trade off against each other. In this example, $\lambda(p) = 1-p^2$, satisfying Assumptions 1–3.

to deliberate fraud, such as when data is misreported or fabricated.[11]

I make a number of assumptions on the way speed and reproducibility trade off against each other, as reflected in the speed function λ.

Assumption 1 (The speed function is decreasing). For all $p, q \in [0, 1]$, if $p < q$, then $\lambda(q) < \lambda(p)$.

Assuming that λ is decreasing means that the scientist expects to take more time to do research that is less likely to be erroneous, by collecting more data or being more thorough, say. One might object that in some situations (such as when the scientist discovers a mistake in her previous work) the scientist's confidence in the reproducibility of her work might go down instead of up over time, seemingly in violation of this assumption. But this misinterprets the function λ. This function gives, for each desired reproducibility p, the scientist's *ex ante* expectation of how long it would

[11]But see Justin P. Bruner, "Policing Epistemic Communities," *Episteme*, x, 4 (December 2013): 403–16; and Liam Kofi Bright, "On Fraud," *Philosophical Studies*, clxxiv, 2 (February 2017): 291–310.

take for her confidence in her result to reach at least level p. So if the scientist's confidence was at p or above before discovering the mistake, she would already have published, but if her confidence was below p she will have to work until it finally reaches p before publishing.

Thus the model does not capture the dynamics of a scientist's expectations about the duration of the project as they change over time. However, I expect that similar conclusions would be reached in a suitable dynamic model by evaluating the scientist's expectations at any given time.

Assumption 2 (The speed function is concave). For every $p, q, t \in [0, 1]$,

$$t\lambda(p) + (1-t)\lambda(q) \leq \lambda(tp + (1-t)q). \tag{1}$$

This assumption reflects a kind of decreasing marginal returns. As the reproducibility p is lowered, the expected speed λ increases ever more slowly: writing the paper itself takes time, which becomes relatively more significant if the scientist spends relatively little time on the research content. Conversely, if the scientist aims for higher reproducibility (increasing p), the speed λ drops off ever faster. More time is required to increase p from 0.8 to 0.9, say, than from 0.7 to 0.8.[12]

Assumption 3 (No perfect work). $\lim_{p\to 1} \lambda(p) = 0$.

This assumption asserts that the scientist cannot deliver perfect work (in the sense of zero probability of errors), no matter how slowly she works. This reflects the fact that there is no certainty in science: it is always possible for any fact or discovery to be overturned, as Lakatos and Quine have argued.

Note that these assumptions imply the following restrictions on the expected completion time $\mu(p) = 1/\lambda(p)$: the expected completion time is increasing and convex, and diverges to infinity as p approaches one. These restrictions can be given justifications analogous to the above.

[12] This observation goes back at least to Charles Sanders Peirce, "Note on the Theory of the Economy of Research," *Operations Research*, XV, 4 (1967 [1879]): 643–48, at p. 644.

The above assumptions also imply that the expected speed is a continuous function of reproducibility, which may be unrealistic when (say) experimental results arrive in batches, leading to discontinuous jumps in reproducibility. This is only a problem for my model if such discontinuities are sufficiently common and predictable that the scientist can anticipate them (since the speed function reflects her *ex ante* expectations). This requires not only that the scientist knows in advance that experimental results arrive in batches, but also that she can predict fairly accurately what level of reproducibility she will reach with the first batch.

Such cases may arise; my claim here is not to capture the choices of all scientists everywhere, but a large range of cases. The types of cases excluded from the model are those in which evidence is gathered in discrete amounts, with relatively predictable effects on the scientist's confidence in her results, and where the scientist is in a position to decide whether or not to gather more evidence after seeing some initial results.

II. PEER REVIEW, CREDIT, AND SOCIAL VALUE

For reasons I outlined above, I assume the scientist gets credit only for published work. Whether the scientist's work is published is determined through *peer review*. The purpose of peer review is to determine the accuracy of the scientist's work.

Suppose this "pre-screening" works perfectly: all and only those papers that are in fact accurate are accepted. The scientist does not know whether her paper is accurate. She only knows the reproducibility p: her own credence that it is accurate. So from the scientist's perspective, if she produces a paper with reproducibility p, there is a probability p that the journal publishes it. Writing c_a for the average amount of credit that accrues to the scientist for a published accurate result, the scientist's expected credit per unit time is a function C of the chosen reproducibility p given by $C(p) = c_a p\lambda(p)$.

In reality the peer review system cannot perform its "pre-screening" per-

fectly. Some accurate results get rejected (so-called *false negatives*), while some erroneous results get accepted (*false positives*). Following common usage in statistics I write α for the probability of a false positive and β for the probability that a false negative is avoided, or equivalently, that an accurate result is accepted.

I assume that peer review is imperfect in the sense of there being a positive probability of false positives ($\alpha > 0$). I remain agnostic on the possibility of false negatives ($\beta \leq 1$) although it seems reasonable to assume that those occur as well. I do assume that accurate results, like erroneous results, have a non-negligible chance of acceptance ($\beta > 0$).

Assumption 4 (Imperfect peer review). The peer review acceptance probabilities are such that $\alpha > 0$ and $\beta > 0$.

I write c_e for the average amount of credit accrued for a published erroneous result. Research that failed to reproduce is still frequently cited as if it were accurate,[13] even after a formal retraction.[14] In other cases the fact that the proposed hypothesis has fallen out of favor does not prevent it from being a credit-worthy contribution to science, such as with Priestley's work on phlogiston. This suggests that erroneous publications are worth some credit ($c_e > 0$), although I will not assume this: I allow that credit for erroneous publications may sometimes be negative. The point here is simply that erroneous publications can influence a scientist's credit stock.

Putting all of this together yields the following. The scientist works on her study at expected speed $\lambda(p)$. The result is accurate with probability p. In this case it gets published with probability β and this publication is worth c_a units of credit. With probability $1 - p$ the result is erroneous, which

[13] Athina Tatsioni, Nikolaos G. Bonitsis, and John P. A. Ioannidis, "Persistence of Contradicted Claims in the Literature," *Journal of the American Medical Association*, CCXCVIII, 21 (Dec. 5, 2007): 2517–26.

[14] John M. Budd, MaryEllen Sievert, and Tom R. Schultz, "Phenomena of Retraction: Reasons for Retraction and Citations to the Publications," *Journal of the American Medical Association*, CCLXXX, 3 (Jul. 15, 1998): 296–97.

leads to a publication worth c_e units of credit with probability α. Thus the scientist's expected credit per unit time, as a function of p, is

$$C(p) = c_a\beta p\lambda(p) + c_e\alpha(1-p)\lambda(p). \tag{2}$$

To compare the individually optimal (that is, credit-maximizing) tradeoff between speed and reproducibility to the socially optimal tradeoff, it is important to be explicit about what is meant by the *social value* of a research study. Here I have in mind the contribution that it makes to science as a social enterprise, which in turn benefits society. This is reflected in the utilization of the work by other scientists, and the extent to which it or work based on it finds its way into society, in the form of a new medicine, for example.

What is the expected social value V of the scientist's research? I assume that research can have social value only when it is published. The probabilities of publication α and β, the reproducibility p, and the expected speed $\lambda(p)$ are all as above. Hence

$$V(p) = v_a\beta p\lambda(p) + v_e\alpha(1-p)\lambda(p), \tag{3}$$

where v_a is the average social value of an accurate result, and v_e the average social value of an erroneous result. The social value function looks very similar to the credit function, but in section III I argue that there is reason to expect v_e to differ systematically from c_e.

Assumption 5 (Positive value). Accurate results have positive credit value ($c_a > 0$) and social value ($v_a > 0$).

With Assumption 5 in place, the first result follows. It states that the functions C and V have unique maxima: there is a particular reproducibility that a rational credit-maximizing scientist would choose, and there is a particular reproducibility that maximizes the social value of the scientist's

contribution.[15]

Theorem 1 (Unique maxima). If Assumptions 1–5 are satisfied, then there exist unique values $p_C^* < 1$ and $p_V^* < 1$ that maximize the functions C and V respectively, that is,

$$C(p_C^*) = \max_{p\in[0,1]} C(p) \qquad \text{and} \qquad V(p_V^*) = \max_{p\in[0,1]} V(p). \tag{4}$$

Note that $p_V^* < 1$. This means that, even from the social perspective, perfect reproducibility is not a goal worth striving for. In other words, even if the scientist was "high-minded" in the sense that she only cared about maximizing the social value of her scientific work, she should not strive to avoid error at all cost.

This is a more or less direct consequence of the "no perfect work" assumption and hence reflects the insight of Lakatos and Quine that there is no certainty in science. It means that even in a science functioning perfectly, a tradeoff between speed and reproducibility must be made, and hence errors should be expected. I emphasize this point since philosophers of science and epistemologists have said a lot about error avoidance but relatively little about how to achieve this in a reasonable time frame.[16]

III. THE INCENTIVE TO RUSH INTO PRINT

Theorem 1 does not say how the credit-maximizing reproducibility p_C^* and the social-value-maximizing reproducibility p_V^* relate to each other. Establishing such a relation requires further assumptions on the parameter values.

The first assumption is that credit is awarded for (accurate) scientific

[15] A proof is provided in appendix A of Remco Heesen, "Expediting the Flow of Knowledge Versus Rushing into Print," *PhilSci-Archive* (2018), `http://philsci-archive.pitt.edu/15161/`, where this result is labeled Theorem 3.1.

[16] See Michael Friedman, "Truth and Confirmation," this JOURNAL, LXXVI, 7 (July 1979): 361–82; and Remco Heesen, "How Much Evidence Should One Collect?," *Philosophical Studies*, CLXXII, 9 (September 2015): 2299–313.

work proportional to its social value ($v_a = c_a$). Since, for all I have said so far, credit and social value are measured on unspecified interval scales, this may be viewed merely as fixing these scales (without loss of generality). Merton and Strevens argue that there is in fact a substantive link between the credit given for scientific achievements and the social value resulting from them.[17]

How about the social value of an erroneous result v_e? I take it that errors are distracting or actively misleading more often than they are instructive. Take, for instance, a study which erroneously finds that a particular medicine helps cure some disease. Once the erroneous finding is published, it takes more time and effort to set the record straight than it would have in the absence of the erroneous publication. Moreover, before the error is corrected there may be negative consequences for other research and public health.[18]

So it seems to me that erroneous results are, on average, at best socially neutral, if not socially harmful: $v_e \leq 0$. And I suggested above that they may still yield positive credit: $c_e > 0$. However, I need not insist on these conclusions. The weaker assumption that the social value of erroneous results is less than the credit given for them ($v_e < c_e$) suffices for my argument.

Assumption 6 (Credit and social value). Accurate results are awarded credit proportional to their social value ($c_a = v_a$), while the social value of erroneous results is less than the credit given for them ($v_e < c_e$).

The main result of this paper can now be stated. It says that the imperfections in the peer review system and the way credit is awarded systematically favor lower levels of reproducibility. That is, a scientist who maximizes expected credit will set reproducibility no higher than the optimal level from

[17]Robert K. Merton, "Priorities in Scientific Discovery: A Chapter in the Sociology of Science," *American Sociological Review*, XXII, 6 (December 1957): 635–59, at p. 659; and Strevens, "Role of the Priority Rule," *op. cit.*, p. 78.

[18]There may even be negative consequences after the error is corrected. See Budd, Sievert, and Schultz, "Phenomena of Retraction," *op. cit.*; and Tatsioni, Bonitsis, and Ioannidis, "Persistence of Contradicted Claims," *op. cit.*

the perspective of maximizing social value.[19]

Theorem 2 (Rushing into print). Let Assumptions 1–6 be satisfied, and define p_C^* and p_V^* as in Theorem 1. Then $p_C^* \leq p_V^*$.

Given the assumptions, there is a credit incentive to produce research at a systematically lower reproducibility than is socially optimal. This result depends crucially on the imperfections in the peer review system, and in particular the possibility of false positives: if $\alpha = 0$ and $\beta > 0$ then Assumptions 1–3 and 5 are sufficient to show that the functions C and V have unique maxima, and that these maxima are equal. Given imperfect peer review, it makes sense for scientists to produce lots of papers and "see what sticks" rather than spend too much time perfecting a paper, and since any resulting errors hurt society more than the scientist, the result follows.

I now briefly consider two objections. First, for all Theorem 2 says it could be that $p_C^* = p_V^*$, the happy case in which individual incentives and social optimality align exactly. However, this happens only if either the value of erroneous results is so high that it is socially optimal to have no concern for reproducibility ($p_C^* = p_V^* = 0$, not a particularly happy case) or the speed function is not differentiable at the point of optimality. If these two situations are ruled out the inequality is strict ($p_C^* < p_V^*$).[20]

Second, one may object to the reproducibility p being a subjective probability. While the scientist may estimate credit subjectively, what matters from the perspective of social value is the objective reproducibility of her work. I reply that an important aspect of a scientist's training is learning to assess evidence objectively, so the scientist's subjective reproducibility should be close to the objective one.[21] Insofar as they differ, the scientist is more

[19] This result is proven as Theorem 3.2 in Heesen, "Expediting Versus Rushing," *op. cit.*, appendix A.

[20] Heesen, "Expediting Versus Rushing," *op. cit.*, Theorem 3.3.

[21] For some evidence that scientists as a group are good at estimating reproducibility in advance, see Camerer et al., "Replicability of Social Science Experiments," *op. cit.*

likely to be overconfident than underconfident. My result still holds if the objective reproducibility is less than or equal to the subjective probability.[22]

What does this result mean for real scientists, who may care about other things than maximizing credit, and who may be less than fully rational? Insofar as credit acts as a selection mechanism in science, this means scientists who rush into print are more likely to succeed than scientists who do not, so one should expect rushing into print to increase over time.[23] Thus there is a *structural* misalignment of incentives, the effect of which is to push scientists in the direction of rushing into print.

I think this misalignment is worth addressing, but one might object that there could be countervailing motivations (goals of scientists other than credit) or systematic irrationalities that make scientists choose socially optimal reproducibility levels despite my argument. It would be a surprising coincidence if other motivations or irrationalities balanced out the incentive to rush into print exactly, but I do not have an argument to rule this out. The objection does illustrate the more general point that in evaluating potential policy responses we should consider not just their effect on the issue at hand (here, the credit incentive to rush into print) but also what the potential side effects might be (here, effects on other motivations or irrationalities) and how they can be managed. This is one reason why I stop short of recommending any particular action in the next section.

IV. WHAT CAN BE DONE?

What can be done about this misalignment of incentives? One solution is to eliminate imperfections in the peer review system. Without those imperfections credit incentives are perfectly aligned with the social optimum in my model. But this is a lot to ask: it requires reviewers at scientific journals to

[22]Heesen, "Expediting Versus Rushing," *op. cit.*, Corollaries 3.1 and 3.2.

[23]See also Paul E. Smaldino and Richard McElreath, "The Natural Selection of Bad Science," *Royal Society Open Science*, III, 9 (Sep. 21, 2016): 160384.

be perfect predictors of whether a study will be successfully reproduced.

However, I noted that the misalignment of incentives in the model is exclusively caused by false positives. So reducing those can bring the credit-maximizing optimum closer to the social optimum. This seems to recommend conservative editorial practices: rejecting papers even based on fairly minimal doubts about their reproducibility. But if reducing false positives also leads to more false negatives the effect will be that the maximum social value is itself lowered, even if the credit-maximizing optimum is brought closer to it. Investigating this further tradeoff is beyond the scope of this paper.

A different way to eliminate imperfections in the peer review system involves getting rid of peer review altogether (possibly replacing it with post-publication peer review). But even such a drastic rethinking of the way scientific research is disseminated would not avoid this problem. The problem arises because scientific work needs to be evaluated in some way in the short run (scientists need to decide what to read and what to cite, for example). Hence the existence of peer review in its current form is not essential to the incentive to rush into print.

Another solution focuses on the amount of credit given for irreproducible results. If the credit given to irreproducible results matched the social value of those results more closely, the gap between the credit-maximizing optimum and the social optimum would be reduced. It would help if there were broader general awareness of which research has been refuted, but this may be hard to achieve. More specifically, one might aim to make hiring and promotion committees more aware of candidates' refuted results.

A third solution aims to compensate for the misalignment. For example, Nelson, Simmons, and Simonsohn have suggested limiting the number of papers scientists may publish per unit time.[24] But the limit on the number of papers would have to be just right to balance out the incentive to favor

[24]Leif D. Nelson, Joseph P. Simmons, and Uri Simonsohn, "Let's Publish *Fewer* Papers," *Psychological Inquiry*, XXIII, 3 (2012): 291–93.

speed over reproducibility without overshooting the optimum in the other direction. This problem is exacerbated by the fact that different scientists may have different speed functions, which may require different publication limits to create the best incentive structure.

In this paper I have focused on rushing into print, without denying that publication bias, data dredging, and other factors may also contribute to reproducibility problems. But whereas the latter wear their corresponding solutions on their sleeve (negative results should be publishable, scientists should commit to pre-analysis plans), this discussion suggests that the solution to rushing into print is much less clear, if one exists at all. On this issue, at least, it seems that the reward structure of science does not incentivize socially beneficial choices.

REMCO HEESEN

University of Western Australia