Why the Reward Structure of Science Makes Reproducibility Problems Inevitable*

Remco Heesen†

September 20, 2017

Abstract

Recent empirical work has shown that many scientific results may not be reproducible. By itself, this does not entail that there is a problem (or "crisis"). However, I argue that there is a problem: the reward structure of science incentivizes scientists to focus on speed and impact at the expense of the reproducibility of their work. I illustrate this using a well-known failure of reproducibility: Fleischmann and Pons' work on cold fusion. I then 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. In the conclusion I consider possible solutions and implications for how Fleischmann and Pons' work should be evaluated.

*Thanks to Kevin Zollman, Michael Strevens, Stephan Hartmann, Teddy Seidenfeld, Jan Sprenger, Liam Bright, Cailin O'Connor, Seamus Bradley, Conor Mayo-Wilson, and audiences at the University of Tilburg, 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.

†Faculty of Philosophy, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, UK. Email: rdh51@cam.ac.uk.

1 Introduction

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.[1]

Reproducibility has come under increased scrutiny, especially in the fields of medicine and psychology. There has long been "a general impression that many results that are published are hard to reproduce" (Prinz et al. 2011), which has recently begun to be empirically tested. Two studies by pharmaceutical companies reported that they could reproduce less than a quarter of results in cancer biology (Prinz et al. 2011, Begley and Ellis 2012). A large systematic study of results published in prominent psychology journals found that less than half could be reproduced (Open Science Collaboration 2015). A similar study for cancer biology is currently underway (Nosek and Errington 2017).

This has led to talk of *reproducibility problems* (or a "crisis"): the above results are taken to show that a smaller number of scientific results is reproducible than would be "expected" if science was being done "properly". However, empirically measured reproducibility rates cannot prove this by themselves as it is not obvious what reproducibility rates should be expected. The following toy model illustrates this (based on Ioannidis 2005).

Consider a scientific community engaging in various projects with the aim to make major scientific discoveries. Suppose that only one in every twenty-one projects leads to a genuine discovery, but there is a false discovery rate $\alpha = 0.05$. Then, on average, two out of twenty-one projects will claim a discovery, but only half of these will be genuine.

[1] Determining whether a particular result should count as "the same" can be a complicated affair. The field of statistics is largely devoted to this task.

In such a world, finding that approximately half of all scientific discoveries can be reproduced would in fact be expected if science was being done properly. More generally, the unknown number of genuine scientific results waiting to be discovered (relative to the total number of scientific projects) influences the reproducibility rates that can be expected. So the empirical reproducibility rates cited above do not suffice to show that there is a problem.

The aim of this paper is to show that there is in fact a problem, and to diagnose one source of the problem. The diagnosis is that due to the *reward structure of science* scientists have an incentive to produce research that is less likely to be reproducible than we should want. This diagnosis is more general than those that identify particular statistical practices or journal practices as the source of the problem, implying that the problem is harder to solve than some have thought.

The reward structure of science is centered around *credit*. Credit is acquired by receiving recognition for one's work, e.g., by having it published in a scientific journal, by having it cited, or by receiving a prestigious award for it. Because their careers may depend on it, scientists are very concerned about credit. This point has long been recognized by philosophers of science like Hull (1988, chapter 8), Kitcher (1993, chapter 8), Strevens (2003), and Zollman (forthcoming) and sociologists such as Merton (1957, 1969) and Latour and Woolgar (1986, chapter 5).

This raises the question what behaviors scientists are likely to engage in to get credit. Philosophers and economists have used rational choice models to answer this question. Kitcher (1993, chapter 8) and Strevens (2003) argue that credit can incentivize scientists to distribute themselves over research programs in a way that is closer to optimal than if they were individually epistemically rational. Dasgupta and David (1994) and Zollman (forthcoming) argue that credit incentives speed up the progress of science. Boyer-Kassem and Imbert (2015) argue that credit incentives encourage collaboration be-

tween scientists. And Boyer (2014), Strevens (2017), and Heesen (2017a) argue that credit incentives can motivate scientists to share their work widely.

This literature largely focuses on the positive effects of credit incentives.[2] In contrast, it has been suggested that the incentive structure of science may be (at least partially) responsible for reproducibility problems—a negative effect. I distinguish four such claims.

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.*

Claim (Publication bias)**.** *Scientists are incentivized to favor positive results at the expense of reproducibility.*

Claim (Novelty bias)**.** *Scientists are incentivized to favor novel results at the expense of reproducibility.*

Claim (Checking bias)**.** *Scientists are disincentivized to attempt to reproduce each other's work, reducing the incentive to make sure their own work is reproducible.*

Publication bias, novelty bias, and checking bias each pick out relatively specific features of the reward structure of science.[3] Publication bias originates from journals' preference for positive results (Easterbrook et al. 1991, Egger and Smith 1998).[4] A "positive" result usually means that a statistical

[2]Largely, but not exclusively. For example, it has been argued that credit incentives contribute to herding behavior (Strevens 2013) and a productivity gap between male and female scientists (Bright 2017b).

[3]I do not mean to take a strong stand here on whether these phenomena are appropriately called biases. The name "publication bias" is well established in the literature and I am simply naming novelty bias and checking bias by analogy to this existing terminology.

[4]I focus on journals here because journal publications are an important way to be rewarded for scientific work. This includes both a preference by editors and reviewers for positive results and a preference by scientists to only submit positive results in anticipation of such a preference (note that the latter can happen even if journals do not actually have such a preference).

hypothesis that a certain experimental condition has "no effect" (a null hypothesis) is rejected. As a result, evidence favoring a null hypothesis is not published, biasing the scientific literature (the so-called "file drawer problem" Rosenthal 1979).

Novelty bias and checking bias originate from journals' preference for novel results. In particular, journals are generally not interested in publishing *direct replications*, i.e., studies that follow (as much as possible) the same experimental design as a previously published study (Neuliep and Crandall 1990). This incentivizes scientists to skew (the presentation of) their results to focus on novel findings (novelty bias). It also means scientists do not expect that anyone will attempt to reproduce their work, weakening their incentive to make sure their work is reproducible (checking bias).

Publication bias, novelty bias, and checking bias each suggest their own solution. Publication bias would be prevented or seriously reduced if null results were regularly published and rewarded (Ioannidis 2006). Novelty bias and checking bias would be seriously reduced if replication studies were regularly published and rewarded (Koole and Lakens 2012).

The incentive to rush into print, on which I focus in this paper, is different in two ways. First, as I will argue, it does not depend on fairly specific features of the reward structure of science, but rather on the general facts that scientific work is rewarded and that these rewards are determined at least partially before it is known whether the work is reproducible. Second (and as a consequence of the first), it is much less obvious how the incentive to rush can be reduced or eliminated.

Rushing into print describes the incentive that scientists have to focus on the speed with which they can produce results and/or the impact of those results, and the corresponding lack of focus on the reproducibility of these results. This is usually described with the phrase "pressure to publish" (Fanelli 2010, Prinz et al. 2011), although I argue the phenomenon is not essentially tied to journal publications.

I illustrate rushing into print in section 2 with a case study in which a concern for credit led to the publication of research that other scientists were then unable to reproduce: Fleischmann and Pons' work on cold nuclear fusion.

I then provide a rational choice model of scientists' decision how much time to spend on a particular study before trying to publish it and I compare the credit-maximizing choice to the choice that is optimal from a social perspective. The model has three basic ingredients. First, the fact that speed and reproducibility trade off against each other (section 3). Second, the fact that scientists get rewarded for publications. And third, the fact that the system of peer review fails to predict perfectly which work will be successfully reproduced.

The point of the rational choice model is to show that scientists have an incentive to produce work that is less likely to be reproducible than is socially desirable, i.e., an incentive to rush into print (section 4). As I argued above, this is not something that can be shown by empirical studies of reproducibility rates. I also show that the particular journal preferences responsible for publication bias, novelty bias, and checking bias are not necessary for reproducibility problems to arise. They may, however, exacerbate such problems.

In the conclusion (section 6) I summarize my results. I also discuss possible ways to disincentivize rushing into print. And finally, I discuss the extent to which my results support an interpretation of Fleischmann and Pons' cold fusion research as a case of rushing into print.

My focus in this paper is exclusively on the incentives that scientists have to do their work in a way that is less likely to be reproducible. This *incentive-based* approach to reproducibility problems may be contrasted with a *methodological* approach. A methodological approach focuses on identifying practices that scientists engage in that may lead to irreproducible research.

For example, various choices have to be made in conducting a research

study: Should outliers be excluded? What statistical test should be used? And so on. These choices have been called "researcher degrees of freedom" (Simmons et al. 2011). If scientists run their analysis multiple times (varying how these choices are made) and report only cherry-picked[5] results, this creates a biased publication record (Ioannidis 2005, Simmons et al. 2011).

Focusing on methodology invites different suggestions for improving reproducibility than focusing on incentives: e.g., pre-registration of studies and more complete reporting of results. However, such suggestions will not eliminate researcher degrees of freedom entirely. It is both impossible and undesirable to reduce the scientific method to a fixed mechanical protocol, as Feyerabend (1975) famously argued.

This suggests that if a scientist has an incentive to produce biased results, it will always be possible for her to do so without straying from the methodological norms of her field. For this reason I think methodological suggestions for improving reproducibility are unlikely to be as effective as their proponents hope unless the incentives leading to irreproducible research are also addressed.

2 Cold Fusion

In this section I use a case study to argue that the pressure to publish can lead to the publication of research that cannot be reproduced. The rest of the paper aims to show that this is a structural rather than an incidental problem.

On March 23, 1989, two established and respected chemists named Martin Fleischmann and Stanley Pons gave a remarkable press conference at the

[5]The discussion above has suggested what kind of cherry-picking scientists may engage in. For example, results that reject a null hypothesis would be favored over those that fail to reject a null, results that have a "novelty" factor would be favored, and so on. This can be done consciously (for careerist reasons) or unconsciously ("this result fails to support our hypothesis, so something must have gone wrong").

University of Utah. They claimed that by loading a palladium rod with deuterium through electrolysis, they had turned the rod into a source of energy, producing up to four times as much heat as they put in.

They hypothesized that the deuterium atoms might be packed together so closely within the palladium as to force pairs of them together in an energy-producing process known as *nuclear fusion*. Conventional wisdom held that a sustained, controlled fusion reaction—the kind needed for a viable source of energy—requires temperatures over a hundred million degrees. Now two chemists claimed to be able to achieve the same thing at room temperature, hence the name *cold fusion*.

A media hype ensued, as cold fusion held the promise of a clean and nearly boundless source of energy. Given these implications, and Fleischmann and Pons' impeccable credentials as experimentalists, scientists around the world dropped what they were doing to attempt to reproduce the experiment.

Shortly after the press conference, a number of announcements were made by researchers seeing similar phenomena. But as time passed their claims came under heavy criticism. The excess heat measurements were attributed to mistakes in accounting for the potential recombination of gases released during the experiment. The neutron measurements (Fleischmann and Pons' other important piece of evidence) could not be replicated with more sophisticated equipment. After the meeting of the American Physical Society (APS) in May 1989, the tide shifted from a mixture of excitement and skepticism to a consensus that Fleischmann and Pons had been mistaken; the phenomenon was deemed irreproducible.

The current scientific consensus is that it is not possible to achieve cold fusion at meaningful rates. In their books on the case, Close and Huizenga judge that Fleischmann and Pons "went public too soon with immature results" (Close 1991, p. 328) and that their "gamble to go public... is the scientific fiasco of the century" (Huizenga 1993, p. 214). What led Fleischmann and Pons to make this fateful decision to "go public"?

At the nearby Brigham Young University, physics professor Steven Jones and his team had been working on a very similar project. The main differences were that Jones was primarily interested in explaining the heat at the center of the Earth (rather than creating a new source of energy) and that he focused on measuring neutron production rather than excess heat.

The two teams first became aware of each other in September of 1988 and interacted a number of times. In February of 1989, Jones announced that he was going to present his data at the APS meeting in May and was planning to submit an article to a journal soon. Fleischmann and Pons were not ready for this yet. They were confident that some experiments produced excess heat, but much remained to be investigated. They indicated that they wanted to do another eighteen months of research before going public (Pool 1989, Huizenga 1993, p. 18).

The two groups agreed to a compromise: they would submit their results to *Nature* simultaneously on March 24. Jones has claimed that there was a further agreement not to publicize the work until that time, but Pons has denied this (Pool 1989, Close 1991, p. 94). Either way, Fleischmann and Pons did publicize their work: they sent a manuscript to the *Journal of Electroanalytical Chemistry* on March 11, and held the infamous press conference on March 23.

Their goal in going public was to establish priority—and thus claim credit—for the cold fusion research, especially relative to Jones (Huizenga 1993, p. 19). This may seem unnecessary in hindsight, as Jones' experimental results were quite different (measuring neutrons rather than heat) and of such a different order of magnitude as to hold no promise for a viable source of energy, thus posing no threat to the originality or importance of Fleischmann and Pons' work. But this was not so clear at the time, as Fleischmann explained later.

> We could not tell whether Jones had heat data or was planning to look for this. How could one tell? He was certainly thinking about

> fusion as a source of heat in the Earth. If he was going to say *that* in the paper, which was surely his intention to do, it would almost certainly destroy any possibility of patent protection (quoted in Close 1991, pp. 99–100).

Thus, both the decision to agree to publish simultaneously with Jones and the later decisions to submit to a different journal before Jones and hold a press conference were made out of a concern for credit. Fleischmann and Pons were aware that their results were still preliminary but went public anyway.

So a concern for credit led to the publication of research that other scientists were unable to reproduce. The model presented next aims to establish that this is a structural problem: scientists have a credit incentive to rush into print. In doing so, the model also lends support to the claim that Fleischmann and Pons did nothing irrational by going public when they did, despite Close and Huizenga's criticism of this decision (as I suggest in section 6).

3 A Tradeoff Between Speed and Reproducibility

Consider a scientist—or a team of scientists, such as Fleischmann and Pons—working on a research study. When should she attempt to publish her work? As the case of Fleischmann and Pons illustrates, getting credit for the work is an important consideration.

As I mentioned in the introduction, scientists' concern for credit is well-documented and understandable, given its importance to scientific careers (Merton 1957, 1969). Because I am interested in what scientists have a credit incentive to do, here I make the methodological assumption that credit is their *only* concern.

Since the scientist aims to maximize the amount of credit she accrues per unit time, she prefers to work faster rather than slower (all else being equal): the concern for credit entails a concern for *speed* (to be defined more formally below). At the same time, working faster reduces *reproducibility*. By reproducibility I mean, loosely speaking, the likelihood that the result of the research study (e.g., "cold fusion is a viable source of energy") is reproduced if someone attempts to do so.

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, i.e., 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, i.e., 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.[6] This definition should be interpreted broadly, applying to both experimental and non-experimental contributions (e.g., 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 (think of the eighteen more months of research

[6]Note that it follows from these definitions that reproducibility goes up if fewer attempts to reproduce scientific results are made, because results that are never tested count as accurate. For this reason, the present model is not suitable to study strategic decisions scientists might make regarding whether to reproduce others' work. See Bruner (2013) for a model in which the incentives to reproduce others' work are considered.

Fleischmann and Pons wanted to do). This is reflected in the model by the *speed function* λ. The value $\lambda(p)$ reflects the speed at which the scientist works if the desired reproducibility is p (see figure 1). More specifically, $1/\lambda(p)$ is the expected time until completion of the study (so $\lambda(p)$ is the number of studies "like this one" that the scientist would expect to complete per unit time). This reflects the scientist's *ex ante* expectation about the duration of the study.

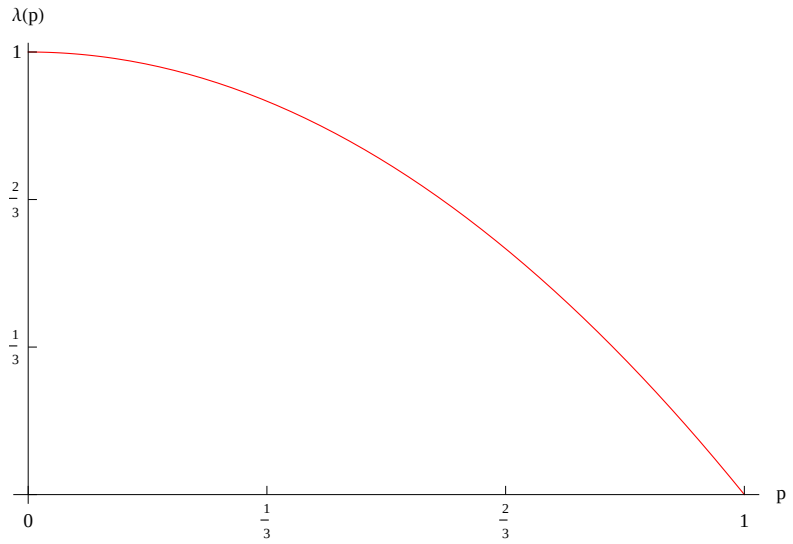

Figure 1: p and λ trade off against each other. In this example, $\lambda(p) = 1-p^2$, satisfying assumptions 1–3.

Reducing reproducibility (lowering p) allows the scientist to publish faster. "Rushing" the work in this way may cause mistakes to be made that make it irreproducible. The present model is not intended to investigate incentives related to deliberate fraud, such as when data is misreported or fabricated (but see Bruner 2013, Bright 2017a).

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 it takes more time to do research that is less likely to be erroneous, e.g., by collecting more data.

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

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

This assumption reflects a kind of decreasing marginal returns. As the reproducibility p is lowered, the speed λ increases ever slower: 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, e.g., to increase p from 0.8 to 0.9 than from 0.7 to 0.8 (Peirce 1967 [1879]).

To put the point in more statistical terms: the size of the standard error of a parameter estimate gives an indication of the reproducibility of the result. As n (the amount of data collected) increases, this width decreases at a rate proportional to $\sqrt{n}$. So there are decreasing marginal returns from investing the time to collect more data.

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 (1978) and Quine (1951, section VI) have argued.

The above assumptions imply that the speed function is continuous, which may be unrealistic when (say) experimental results arrive in batches, leading to discontinuous jumps in p. 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). Such cases may arise; my claim here is not to capture all scientists everywhere, but a large range of cases.

4 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 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 perfectly predict the success of future replications. Some accurate results get rejected (so-called *false negatives*), while some erroneous results get accepted (*false positives*). An example of the latter is Fleischmann and Pons' paper in the *Journal of Electroanalytical Chemistry*: it passed peer review but was thoroughly discredited within a year of publication (thus satisfying my definition of erroneous). Following common usage in statistics I write α for the probability of a false positive and β for the "power" (the probability that a false negative is avoided, i.e., that an accurate result is accepted).

I assume that peer review is imperfect in the sense of 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 some 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 for a published erroneous result. While such results are eventually "discredited", this does not necessarily equate to zero credit. Research that could not be reproduced frequently still gets cited as if it was accurate (Tatsioni et al. 2007), even after a formal retraction (Budd et al. 1998). In other cases the fact that the proposed hypothesis has fallen out of favor does not prevent it from being a creditworthy contribution to science, e.g., Priestley's work on phlogiston. This suggests that erroneous publications are worth some credit ($c_e > 0$).[7]

Putting all of this together yields the following. The scientist works on her study at 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 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

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

To compare the individually optimal (i.e., 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

[7]On the other hand, some discredited research can actively harm a scientist's career (more so than publishing nothing at all would have done), suggesting that $c_e < 0$. These are usually cases of fraud rather than honest mistakes and so they are not my primary concern here. However, the point here is not to argue that c_e is necessarily positive, but that erroneous publications can influence a scientist's credit stock.

social enterprise, which in turn benefits society. This is reflected by the utilization of the work by other scientists, and by the extent to which it or work based on it finds its way into society, e.g., in the form of a new medicine.

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 speed $\lambda(p)$ are all as above.[8] Hence

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

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 I argue below 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 ($v_a > 0$).*

The first result states that the functions C and V have unique maxima, i.e., 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.

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, i.e.,*

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

[8] Hence the social value V of the scientist's research is more precisely the scientist's own subjective estimate of the expected social value of the research (because p is a subjective probability). This may seem problematic when I use the function V below to argue that credit incentivizes scientists to make choices that are not socially optimal. I address this point in section 5.

Proofs of this and subsequent results are provided in a technical report (Heesen 2017b).

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.

Theorem 1 does not say, however, how the credit-maximizing reproducibility p_C^* or 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 work proportional to its social value, i.e., $v_a = c_a$. 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 (1957, p. 659) and Strevens (2003, p. 78) argue that there is in fact a substantive link between the credit given for scientific achievements and the social value resulting from them.

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 (and perhaps after as well, see Budd et al. 1998 and Tatsioni et al. 2007) there may be negative consequences for other research and public health.

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 choose reproducibility to be no higher than the optimal level from the perspective of maximizing social value.

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. Intuitively, given imperfect peer review it makes sense for scientists to quickly produce lots of papers and "see what sticks" rather than spending too much time perfecting any one paper.

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? It means that whenever they face a research situation that satisfies the assumptions of my model (which I have argued to apply to typical cases of scientific research, including the one Fleischmann and Pons found themselves in) they either rush into print or they could have improved their expected credit if

they had rushed into print. 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, and one should expect rushing into print to increase over time. Thus there is a *structural* misalignment of incentives. This misalignment is worth addressing even if it may not currently actively cause reproducibility problems (e.g., due to countervailing motivations or irrationality).

5 Three Extensions

This section briefly considers three possible extensions of the model, each motivated by a potential objection to my analysis in the previous section.

The first objection points out that theorem 2 leaves open the possibility that $p_C^* = p_V^*$, the happy case in which individual and social incentives 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 actually a very happy case) or the speed function is not differentiable at the point of optimality (see figure 2).

I take these two cases to be highly exceptional. If they are ruled out, a strict inequality can be shown to hold.

Assumption 7 (Limited social value of errors)**.** *The social value of erroneous results (weighted by the chance of acceptance) is less than half that of accurate results:* $\alpha v_e < \beta v_a/2$.

Assumption 8 (The speed function is differentiable)**.** *The function* λ *is differentiable on the interior of its domain, i.e., for all* $p \in (0, 1)$.

Theorem 3 (Strict inequality)**.** *Let assumptions 1–8 be satisfied, and define* p_C^* *and* p_V^* *as in theorem 1. Then* $p_C^* < p_V^*$.

The second objection is based on the fact that in the model the reproducibility p is a subjective probability. While this is reasonable from the

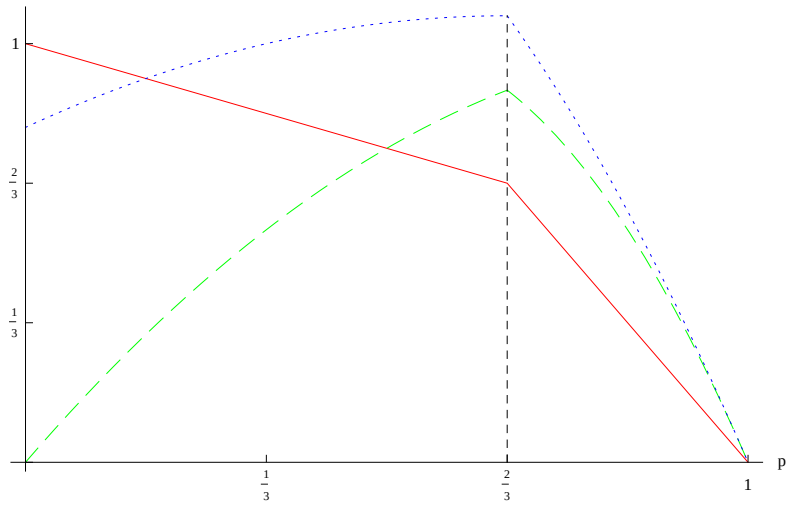

Figure 2: If $\lambda(p) = 1 - p/2$ for $p \leq 2/3$ and $\lambda(p) = 2(1-p)$ for $p > 2/3$ then λ (solid red) is not differentiable at $p = 2/3$. Then the functions C (dotted blue, with $c_a\beta = 2$ and $c_e\alpha = 4/5$) and V (dashed green, with $v_a\beta = 2$ and $v_e\alpha = 0$) may both be maximized there: $p^*_C = p^*_V = 2/3$.

perspective of the scientist's choice of when to go public, it does not seem so reasonable from the perspective of assessing the social value of the scientist's contribution. When it comes to social value, what matters is the actual reproducibility of the result, not the scientist's estimate.

I have two responses to this objection. First, I think it is reasonable to expect the scientist's estimate of reproducibility to be quite good, so that the subjective and the objective probability should be roughly equal. An important part of scientists' training, after all, involves learning how to assess evidence as objectively as possible.

Second, if there is a discrepancy between the scientist's estimate of the reproducibility of the result and its objective reproducibility, the scientist is going to be overconfident more often than underconfident. This is also a result of scientists' training: scientists learn to view the methods they use in their research as the best ones to address the problems they work on (and/or

they self-select into working with the methods they think are best).

To capture this formally, note that the scientist's choice of (subjective) reproducibility determines the objective reproducibility. So I introduce an *objective reproducibility function* o, where $o(p)$ is interpreted as the objective reproducibility that results if the scientist's choice of (subjective) reproducibility is p. Then the foregoing suggests that either $o(p) = p$ or $o(p) < p$.

I also assume that the objective reproducibility function is surjective. This means that any objective reproducibility level is achievable in the sense that there exists a subjective reproducibility level corresponding to it.

Assumption 9 (Confident scientist)**.** *The objective reproducibility function $o : [0, 1] \to [0, 1]$ is surjective, i.e., for all $p \in [0, 1]$ there is a $q \in [0, 1]$ such that $o(q) = p$. Moreover, $o(p) \leq p$ for all $p \in [0, 1]$.*

A credit-maximizing scientist chooses reproducibility p^*_C, the (subjective) probability that maximizes the credit function C. Social value is maximized if the scientist chooses p such that the objective reproducibility $o(p)$ maximizes the social value function V. Given assumption 9, the rushing into print result extends to the case with objective reproducibility.

Corollary 1. *Let assumptions 1–6 and 9 be satisfied, and define p^*_C as in theorem 1. Let q^*_V be any value such that*

$$V(o(q^*_V)) = \max_{p \in [0,1]} V(o(p)).$$

Then $p^*_C \leq q^*_V$.

Corollary 2. *Let assumptions 1–9 be satisfied, define p^*_C as in theorem 1, and q^*_V as in corollary 1. Then $p^*_C < q^*_V$.*

The third objection starts from the observation that the potential *impact* of their work presumably played a role in Fleischmann and Pons' decision to go public. Fleischmann and Pons could perhaps be described as "mavericks", scientists who specialize in high-risk high-reward research in relatively

unexplored areas.[9] This suggests a more general version of the model in which the scientist makes a three-way tradeoff between the desired impact, reproducibility, and speed. The objection suggests that the incentive to rush into print disappears if the scientist's choice whether to go for a high-impact result is taken into account.

Space does not permit a detailed discussion of this version of the model. However, it turns out that under very similar assumptions an analogous result holds: a credit-maximizing scientist chooses a reproducibility level that is structurally lower than what would need to be chosen to maximize social value. See Heesen (2017b) for details.

6 Conclusion

The following four conclusions can be drawn from the work presented in this paper. First, I have argued that under a wide range of plausible conditions scientists have a credit incentive to publish work that is unlikely to be successfully reproduced (relative to the socially optimal reproducibility level). Three key ingredients are responsible for this misalignment of incentives: the tradeoff between speed and reproducibility, the fact that scientists are rewarded for publications, and imperfections in the peer review system.

This misalignment hurts science and society: by definition, any deviation from the social optimum hurts the progress of science and the social benefits of that progress. More specifically, I have shown how credit incentives may contribute to the reproducibility problems that have recently attracted significant attention. They may do so even in the absence of some of the particular phenomena that have previously been identified as culprits (e.g., publication bias, novelty bias, and checking bias).

[9] Close (1991, p. 71) describes Fleischmann in this way. Contrast "followers", scientists mainly making small contributions relatively likely to be accurate. This terminology is used by Weisberg and Muldoon (2009) but the distinction has a long history in philosophy of science (e.g., Hull 1988, p. 474).

What can be done about this? 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 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 (accepting erroneous results for publication). 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 leads to more false negatives (rejecting accurate results) 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 (e.g., scientists need to decide what to read and what to cite). 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 was a 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 would be to try to somehow compensate for the mis-

alignment. For example, Nelson et al. (2012) have suggested limiting the number of papers scientists may publish per unit time. This would create an incentive to favor reproducibility over speed that could in principle balance out the incentive to rush. But the limit on the number of papers would have to be just right to balance out the incentive to favor 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, novelty bias, and checking bias may also contribute to reproducibility problems. But whereas these biases wear their corresponding solutions on their sleeve (scientists should be rewarded for negative results and replications), this discussion suggests that the solution to rushing into print is much less clear, if one exists at all.

The second conclusion is that certainty of reproducibility is neither to be expected nor to be desired. The reason for this is that if scientists were too demanding in perfecting their research before publishing it, nothing would ever get published. The point is hardly new (it goes back at least to Lakatos and Quine), but 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 (cf. Friedman 1979, Heesen 2015), it is worth emphasizing.

The third conclusion concerns next steps in applying this work. One may ask whether there is a way to validate the model. In particular one may want to calibrate the parameter values, for example to see whether there is reason to be worried about rushing into print in a particular case. A good starting point for this kind of work may be in medicine. Here we find relatively well-defined problems (particular diseases) with well-defined solutions (particular treatments). Moreover, measures of the social value of some treatment (e.g.,

survival rates or recovery rates) can be separated relatively cleanly from measures of credit (e.g., citations, prestigious publications, or prizes).

Finally, the work in this paper suggests a reevaluation of Fleischmann and Pons' decision to go public with their work on cold fusion. That decision has been much maligned for being premature. But was it irrational at the time, given the information available to Fleischmann and Pons?

The present work suggest that it may not have been: it can be rational for a credit-maximizing scientist to submit work the reproducibility of which is not yet very firmly established. Fleischmann and Pons were well aware of the uncertainties surrounding cold fusion at the time they went public. They also knew that the risk of being scooped was extremely high. The above considerations suggest (without proving of course) that under these circumstances their decision may have been rational.

Fleischmann and Pons went out on a limb, as every scientist does when she publishes her work. On this occasion, they got burned. But I submit that this was not primarily the result of poor judgment. Rather, they did exactly what other scientists have done on countless occasions: they weighed the risk of going public against the potential reward. That they are now maligned rather than celebrated is largely the result of bad luck.

References

C. Glenn Begley and Lee M. Ellis. Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.

Thomas Boyer. Is a bird in the hand worth two in the bush? Or, whether scientists should publish intermediate results. *Synthese*, 191(1):17–35, 2014.

Thomas Boyer-Kassem and Cyrille Imbert. Scientific collaboration: Do two heads need to be more than twice better than one? *Philosophy of Science*, 82(4):667–688, 2015.

Liam Kofi Bright. On fraud. *Philosophical Studies*, 174(2):291–310, 2017a.

Liam Kofi Bright. Decision theoretic model of the productivity gap. *Erkenntnis*, 82(2):421–442, 2017b.

Justin P. Bruner. Policing epistemic communities. *Episteme*, 10:403–416, 2013.

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*, 280(3):296–297, 1998.

Frank Close. *Too Hot to Handle: The Race for Cold Fusion*. Princeton University Press, 1991.

Partha Dasgupta and Paul A. David. Toward a new economics of science. *Research Policy*, 23(5):487–521, 1994.

P. J. Easterbrook, R. Gopalan, J. A. Berlin, and D. R. Matthews. Publication bias in clinical research. *The Lancet*, 337(8746):867–872, 1991.

Matthias Egger and George Davey Smith. Bias in location and selection of studies. *BMJ*, 316(7124):61–66, 1998.

Daniele Fanelli. Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS ONE*, 5(4), 2010.

Paul Feyerabend. *Against Method*. New Left Books, 1975.

Michael Friedman. Truth and confirmation. *The Journal of Philosophy*, 76 (7):361–382, 1979.

Remco Heesen. How much evidence should one collect? *Philosophical Studies*, 172(9):2299–2313, 2015.

Remco Heesen. Communism and the incentive to share in science. *Philosophy of Science*, 84(4):698–716, 2017a.

Remco Heesen. Analysis and results for 'Why the reward structure of science makes reproducibility problems inevitable'. Technical report, 2017b. URL `https://remcoheesen.files.wordpress.com/2015/03/analysis-and-results.pdf`.

John R. Huizenga. *Cold Fusion: The Scientific Fiasco of the Century*. Oxford University Press, 1993.

David L. Hull. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. University of Chicago Press, 1988.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8), 2005.

John P. A. Ioannidis. Journals should publish all "null" results and should sparingly publish "positive" results. *Cancer Epidemiology Biomarkers & Prevention*, 15(1):185, 2006.

Philip Kitcher. *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford University Press, 1993.

Sander L. Koole and Daniël Lakens. Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, 7(6):608–614, 2012.

Imre Lakatos. *The Methodology of Scientific Research Programmes*. Cambridge University Press, 1978.

Bruno Latour and Steve Woolgar. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, 1986.

Robert K. Merton. Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review*, 22(6):635–659, 1957.

Robert K. Merton. Behavior patterns of scientists. *The American Scholar*, 38(2):197–225, 1969.

Leif D. Nelson, Joseph P. Simmons, and Uri Simonsohn. Let's publish fewer papers. *Psychological Inquiry*, 23(3):291–293, 2012.

James W. Neuliep and Rick Crandall. Editorial bias against replication research. *Journal of Social Behavior and Personality*, 5(4):85–90, 1990.

Brian A. Nosek and Timothy M. Errington. Making sense of replications. *eLife*, 6:e23383, 2017.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015.

Charles Sanders Peirce. Note on the theory of the economy of research. *Operations Research*, 15(4):643–648, 1967 [1879].

Robert Pool. Fusion followup: Confusion abounds. *Science*, 244(4900):27–29, 1989.

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*, 10(9):712, 2011.

W. V. Quine. Two dogmas of empiricism. *The Philosophical Review*, 60(1): 20–43, 1951.

Robert Rosenthal. The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86(3):638–641, 1979.

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*, 22(11):1359–1366, 2011.

Michael Strevens. The role of the priority rule in science. *The Journal of Philosophy*, 100(2):55–79, 2003.

Michael Strevens. Herding and the quest for credit. *Journal of Economic Methodology*, 20(1):19–34, 2013.

Michael Strevens. Scientific sharing: Communism and the social contract. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*, chapter 1. Oxford University Press, 2017.

Athina Tatsioni, Nikolaos G. Bonitsis, and John P. A. Ioannidis. Persistence of contradicted claims in the literature. *Journal of the American Medical Association*, 298(21):2517–2526, 2007.

Michael Weisberg and Ryan Muldoon. Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, 76(2):225–252, 2009.

Kevin J. S. Zollman. The credit economy and the economic rationality of science. *The Journal of Philosophy*, forthcoming.