Where Are The Self-Correcting Mechanisms In Science?

Simine Vazire[1] & Alex O. Holcombe[2]

[1] Melbourne School of Psychological Sciences, University of Melbourne

[2] School of Psychology, The University of Sydney

ABSTRACT

It is often said that science is self-correcting, but the replication crisis suggests that self-correction mechanisms have fallen short. How can we know whether a particular scientific field has effective self-correction mechanisms, that is, whether its findings are credible? The usual processes that supposedly provide mechanisms for scientific self-correction, such as journal-based peer review and institutional committees, have been inadequate. We describe more verifiable indicators of a field's commitment to self-correction. These fall under the broad headings of 1) transparency, which is already the subject of many reform efforts, and 2) critical appraisal, which has received less attention and which we focus on here. Only by obtaining Observable Self-Correction Indicators (OSCIs) can we begin to evaluate the claim that "science is self-correcting." We expect that the veracity of this claim varies across fields and subfields, and suggest that some fields, such as psychology and biomedicine, fall far short of an appropriate level of transparency and, especially, critical appraisal. Fields without robust, verifiable mechanisms for transparency and critical appraisal cannot reasonably be said to be self-correcting, and thus do not warrant the credibility often imputed to science as a whole.

Where Are The Self-Correcting Mechanisms In Science?

"When we observe scientists, we find that they have developed a variety of practices for vetting knowledge – for identifying problems in their theories and experiments and attempting to correct them."

Oreskes (2019, p. 64)

"In reality, mechanisms to correct bad science are slow, unreliably enforced, capricious, run with only the barest nod towards formal policy, confer no reward and sometimes punitive elements for a complainant who might use them"

Heathers (2019)

The assertion that "science is self-correcting" is frequently made by people who claim that science is trustworthy. When serious scientific errors come to light, some respond that such errors are part of the normal, healthy process of science: science sometimes fumbles, but it can be counted on to correct itself. Indeed, even the finding that the most prestigious journals in psychology seem to routinely produce many false positives has been explained away as "science correcting itself". In the introduction to the 2016 volume of the *Annual Review of Psychology*, published just six months after the Reproducibility Project: Psychology results (Open Science Collaboration, 2015) were published, in which over half of findings published in top psychology journals could not be replicated, Fiske, Schacter, and Taylor (2016) concluded that there is no crisis:

> First came a few tragic and well-publicized frauds; fortunately, they are rare—though never absent from science conducted by humans—and they were caught. Now the main concern is some well-publicized failures to replicate [...]. All this is normal science, not crisis. A replication failure is not a scientific problem; it is an opportunity to find limiting conditions and contextual effects. Of course studies don't always replicate.

A few years later, in a report titled *Reproducibility and Replicability in Science,* the National Academy of Sciences (2019) study committee[1] also appeared to downplay the replication crisis. The committee wrote that "The advent of new scientific knowledge that displaces or reframes previous knowledge should not be interpreted as weakness in science". While that statement could be justified, the NAS report left open the question of what a breakdown in science *would* look like. How can we tell the difference between an efficient self-correcting system and one that is making too many unforced errors?

As the NAS report's statement highlighted, the mere existence of scientific errors should not tarnish the reputation of science, and we do not know any way to determine an appropriate level of errors. Therefore, we believe that for science to maintain a strong reputation, and for that reputation to be deserved, scientists should articulate some criteria for assessing whether a scientific field is appropriately self-correcting.

Those criteria should be verifiable to outsiders, so that the stakeholders and the public do not need to rely on scientists' word that their field is self-correcting and therefore trustworthy. We will refer to these as Observable Self-Correction Indicators (OSCIs). Self-correction in science covers a wide range of activities that can vary across fields, and may include audits to assess the prevalence of errors, safeguards to reduce the chances of errors occurring, protocols and incentives for detecting and correcting errors, and mechanisms for disseminating corrections and ensuring that downstream conclusions about theory or applications are also corrected.

Previous suggestions that a particular field is self-correcting typically do not go much further than a mere declaration; details regarding the underlying processes and the effectiveness of the system are rarely, if ever, given. For example, in a blog post entitled "Why science is self-correcting" by the cognitive psychologist Art Markman (Markman, 2010), Markman wrote, "There's no point in scientific misconduct; it is always found." He went on to claim that "The field is able to separate the good results from the bad fairly quickly", not including any detail, which we find frustrating.

Knowing a particular field's level of self-correction is crucial to understanding how much we ought to trust that field - or how much credibility to ascribe to the claims made by that scientific community. The degree to which false claims and findings are identified and speedily corrected likely varies across scientific disciplines and subdisciplines. Thus, claiming that all of science is self-correcting, if only parts of it have reasonably effective self-correction mechanisms, may eventually undermine the credibility of even the most robust sciences, as it links their reputation to that of other sciences with lower credibility. For this reason, it is important that the public be able to appropriately calibrate their trust in any given area of science, and to do so they must be able to evaluate the degree to which any given scientific community has effective mechanisms for self-correction. To do this, we must make the self-correcting mechanisms of science visible.

Individual scientists do not typically self-correct on their own

We suggested above that public trust in science is likely linked to the perception that science has efficient self-correcting mechanisms - and that trust in a scientific field is warranted only to the extent that we can identify robust self-correction mechanisms in that field. Where should we look for these self-correcting mechanisms? Both scientists and non-scientists seem to agree that the self-correcting mechanisms in science, and therefore the trustworthiness of science, do not come from individual scientists' honesty and trustworthiness.

The 2019 edition of a Pew survey of US adults (Funk, Hefferon, Kennedy, & Johnson, 2019) reported that 86% of poll respondents have at least "a fair amount" of confidence in scientists to act in the public interest, which is more than for business leaders, the news media, and elected officials. However, only 11-16% of respondents in the same survey agree that scientists "admit and take responsibility for their mistakes" (when asked specifically about medical, environmental, and nutrition scientists). This cynicism seems to be shared by scientists themselves, who arguably have a fairly negative view of their fellow scientists' integrity. In a survey of early- and mid-career scientists with funding from the US National Institutes of Health, 61-76% of scientists reported that other scientists fail to adhere to Mertonian norms such as disinterestedness and openness (Anderson, Martinson, & De Vries, 2007).

While it is difficult or impossible to estimate the prevalence of dishonesty or incompetence in science, it is clear that the number of papers with errors exceeds, likely by multiple orders of magnitude, the number that are corrected, either by retraction or formal correction. For example, by one estimate, approximately half of psychology journal articles published in eight well-respected journals between 1985 and 2013 include at least one report of a statistic and associated degrees of freedom that is inconsistent with the reported *p*-value. In about one in eight papers, this discrepancy was deemed to be a "gross error", in that while the reported *p*-value was significant, the recalculated *p*-value based on the reported degrees of freedom and test statistic was not, or vice versa (Nuijten et al., 2016). Similarly, in an analysis of genetics articles published between 2005 and 2015, a team of researchers found that about 1 in 5 articles had Microsoft Excel files with gene lists that are affected by gene name errors (e.g., gene names that Excel automatically converts to dates; Ziemann et al., 2016).

So why do members of the public, and presumably scientists themselves, continue to trust science despite this justified cynicism about individual scientists' commitment to self-correction? Many historians, sociologists, and philosophers of science have suggested that trust is warranted despite individual scientists' fallibility because self-correction, and the production of knowledge, is a social process rather than the result of individual scientists' humility, rationality, or accuracy (Campbell, 1988; Haack, 2003; Kuhn, 1962; Longino, 1990; Oreskes, 2019; Sztompka, 2007).

According to this view, healthy scientific communities are structured in ways that encourage, or even ensure, self-correction at the group level. The opening quote by Oreskes illustrates a common refrain, echoed in statements by other scholars of science and by scientists themselves, including methodologist Donald Campbell:

> "The resulting dependability of reports (such as it is, and I judge it usually to be high in the physical sciences) comes from a social process rather than from dependence upon the honesty and competence of any single experimenter. Somehow in the social system of science a systematic norm of distrust (Merton's [1973] "organized skepticism") combined with ambitiousness leads people to monitor each other for improved validity. Organized distrust produces trustworthy reports." (Campbell, 1988, p. 324)

The word "somehow" in the above passage glosses over a number of issues. How does the social system of science achieve this? How do we know that this is so? Others have made similar claims with equally little evidence - it seems that the obvious successes of science (or at least of the physical sciences) is taken as incontrovertible evidence that science uses efficient self-corrective mechanisms. This can also be found in the more informal discourse of journal editors who ask that we "let the self-correcting mechanisms of science take their course" (McConnell, 2020), as if everyone knows that these occur, despite a dearth of explanation or documentation.

In some cases, scholars appealing to social processes of verification and correction have articulated what those processes look like:

> "the existence of strong and rigid self-policing, and self-controlling mechanisms built into scientific communities preventing access by

> incompetent or dishonest people, allows it to be taken for granted that those who pass through the checks of gatekeepers can be trusted (of course, again, this requires 'auxiliary trust' in the integrity and reliability of gatekeepers). More concretely this involves the peer review of publications or grant applications, standardized procedures for obtaining academic degrees and titles, doctoral committees, tenure committees, complex and collective procedures for awarding scientific prizes, juries, disciplinary committees guarding the ethos of science: in extreme cases even courts of law." (Sztompka, 2007, p. 217)

Most scholars who believe that self-correction is highly effective might, if pressed, like Sztompka point to journal-based peer review and disciplinary committees charged with investigating research misconduct, but it is hard to know because it is rare for scholars who claim that science is self-correcting to articulate what the self-correcting processes are. A notable exception is Jamieson and colleagues (2019), who outline some concrete standards that make science trustworthy. Their paper identifies journals and university committees as key stakeholders in enforcing and signalling these standards (see Table 1), though they also discuss the role of individual authors and critics in the scientific community.

Another view is that science self-corrects not by explicit correction of individual manuscripts or papers, but rather by researchers ceasing to cite and repeat the claims in work that contains errors. However, we have not seen any evidence that supports this. In fact the available evidence seems to support the opposite, at least in the short and medium term. Serra-Garcia and Gneezy (2021), for instance, found in a study of top psychology, economics, and general interest journals that "papers that fail to replicate... are cited more than those that are replicable" and observed "no significant change in citation trends, even after the publication of the failed replication".

Traditional scientific institutions are ineffective self-correction mechanisms

Journal-based peer review is thought by many to be the most important error-detection mechanism of science. However, numerous studies suggest that journal-based peer review does not catch the majority of errors. A study at the British Medical Journal sent test papers with nine major errors to over six hundred peer reviewers. Reviewers were given different levels of training to assess whether intervention could help, but for all groups, the mean number of errors detected was three or fewer (Schroter et al., 2008).

An enormous number of errors, then, are likely not detected by journal peer review; other mechanisms are needed for a field to claim that they prioritize self-correction, and therefore that their published findings are trustworthy. In theory, journals themselves provide one potential avenue for such post-publication corrections, via published commentaries, critiques, and corrections. However, anecdotal reports suggest that the bar for publishing reports of errors in a scientific paper can be very high (Heathers, 2015; Pickett, 2020; Friedman & colleagues, in press; Goldacre et al., 2020).

University and publisher- or journal-based committees charged with investigating potential research misconduct are notorious for being unresponsive to evidence, which was one reason the Office of Scientific Integrity, now the Office for Research Integrity, was created in the United

States in 1989 (Gustson, 1994). Journals and universities have strong disincentives to admit that their own research is faulty, and the Office of Research Integrity is woefully under-resourced (Retraction Watch, 2014). Heathers (2015), Pickett (2020), and Friedman and colleagues (in press) have provided first-person accounts of the difficulty of bringing errors or misconduct to light through such formal channels as journals or university disciplinary committees. Clearly, these safeguards often are not effective at ensuring that self-correction is prioritized.

Peer-reviewed journals and traditional university and scholarly institutions evidently cannot be counted upon to reliably detect and correct errors in science. Perhaps they can be reformed (see "Pre-publication peer review" section below), but there are also other avenues now open, thanks to advances in computing and networking. These examples, which we describe in greater detail in the sections below, illustrate the potential for more robust and transparent mechanisms for self-correction than the opaque, conflicted, and highly time-consuming journal peer review processes and misconduct investigation processes.

In short, public trust in science does not seem to be based on trust in individual scientists' commitment to self-correction, but instead on a belief that scientific communities are structured in such a way that self-correction operates at the group level, and is prioritized by the group. However, the few institutions that are typically identified as sources of self-correction in science - namely, peer reviewed journals and university misconduct committees - have not delivered. Nevertheless, so far the public seems to be happy to believe that these social mechanisms for self-correction exist and are effective. However, to maintain public trust in science, and to allow the public to calibrate their trust in different scientific communities appropriately, according to the presence and strength of these mechanisms, it is important to identify these mechanisms and make them visible and verifiable to those outside of a particular scientific community. While confidence in science has remained high during the social, behavioral, and health sciences' replication crisis, reassuring the public with appeals to invisible self-correction mechanisms (or by hand-waving at journal peer review or misconduct committees) may not be effective forever.

Ideally, self-correction mechanisms in science should be visible, quantifiable, and easy for outsiders to evaluate. Below we propose a preliminary list of such Observable Self-Correction Indicators (OSCIs) that are potentially measurable and characteristics of scientific communities rather than individuals.

Observable Self-Correction Indicators (OSCIs)

Self-correction in science requires that work be out in the open (transparency) and actively checked (critical appraisal). In their paper "Signaling the Trustworthiness of Science", Jamieson and colleagues (2019) discuss both transparency-related norms and a "culture of critique" as fundamental to maintaining the trustworthiness of science. Similarly, we have split the OSCIs into two categories: transparency and critical appraisal. While this distinction is not without problems, we believe it is useful. In our experience, many scientists have accepted the value of transparency, but stop there. This is problematic because while transparency is necessary for credibility, it is not enough. Even the most transparently-reported science can be very flawed; the flaws will just be out in the open, without necessarily being detected. Transparency is certainly better than secrecy, but self-correction cannot happen if that transparency is not then used to detect and correct errors (Gelman, 2017; Vazire, 2020).

We have noticed that some researchers, even those who endorse transparency generally, can balk at the idea that other researchers may use their transparency against them, to point out errors or flaws. We believe, however, that the ability to point out errors, or the lack of them, is what transparency is for, and why transparent scientific reports are more credible than opaque reports - it is because the scientist engaging in transparency is giving her critics ammunition that we trust her research more. A scientific community that encourages transparency while discouraging criticism is not one that prioritizes self-correction. Transparency enables critical appraisal, and what the philosopher of science Helen Longino (1990) calls "transformative interrogation", but transparency alone is not enough to ensure it happens.

TABLE 1. Observable Self-Correction Indicators (OSCIs)

Indicators that a field is transparent
Open data
Open code
Open materials and methods
Contributorship
Transparent peer review
Pre-registration/pre-analysis plans
Registered Reports
Level playing field/no barriers to entry
Indicators of critical appraisal in a field
Automated error detection
Computational reproducibility checks
Empirical replicability checks
Pre-publication peer review

Post-publication peer review
Detecting errors and biases in a literature
Strong theory and predictions
Diversity

The importance of transparency, including the facets listed in Table 1, has been explained by others (Asendorpf et al., 2013; Christensen, 2020; Nosek et al., 2018). Because of the extensive amount of attention that has been devoted to transparency, many may assume that we are well on our way to adequate levels of transparency across the sciences. However, it is important to recognize that transparency remains very uneven across fields (e.g., Tenney et al., 2020). As part of the Reproducibility Project cancer biology, Errington and colleagues (2014) set out to examine the methods of 51 papers which reported the results of 197 experiments. Errington (2019) found that for none of those 197 experiments were they able to design a full protocol based purely on the paper, without communicating with the authors. We are optimistic that the situation is not so dire in most fields, but clearly there is much room for improvement. Fortunately, metascientific efforts to evaluate levels of transparency in specific scientific disciplines have continued (e.g., Bakker et al., 2020; Chin, Pickett, Vazire, & Holcombe, in press; Christensen et al., 2020; Franco, Malhotra, & Simonovits, 2015; Fraser et al., 2018; John, Loewenstein, & Prelec, 2012; Makel et al., 2021; O'Boyle et al., 2017).

Because there is already an extensive literature on the value of transparent and open practices in science, we won't discuss those OSCIs here beyond listing some in the top half of Table 1. Instead, we focus on an aspect of self-correction that has received comparably little attention, our second pillar (bottom half of Table 1): critical appraisal.

Critical Appraisal OSCIs

Critical appraisal is characterized by Longino (1990) as the "collective give-and-take of critical discussion" (p. 79) through which objectivity in science is achieved. Objectivity thus depends on the "depth and scope of the transformative interrogation that occurs in any given scientific community" (p. 79). As one critic of Longino has pointed out,

> "Longino envisions 'critical' discussion and 'critically achieved consensus' (1990, 70, 74). The problem, however, is that she gives us little account of the standards that are to govern these discussions. What will make them sufficiently "critical" to confer objectivity on their conclusions?" (Smith, 2004, p. 145)

The present section is a preliminary attempt at an account of the standards, or indicators, by which a scientific community can be evaluated on their success at achieving critical appraisal. We offer these as a tentative framework for developing an OSCIs program further (together with

transparency OSCIs). We discuss each item on the critical appraisal list in varying levels of depth. The discussion of some is very brief because we do not know of much evidence regarding whether and how they could work as OSCIs. Thus, this list is meant as a work in progress, to be expanded, revised, and adapted as new metascientific evidence, and new ideas, develop.

Automated error checking

One gauge of a field's commitment to self-correction is how many easily-detectable and easily-preventable errors appear in the published record. New tools and algorithms for automated error-checking have emerged that authors can use to prevent some such errors. The Statcheck (Nuijten et al., 2014) and GRIM (Brown & Heathers, 2017) algorithms are two examples. They check whether certain kinds of numbers in a paper are consistent with each other. These tools can be used both to catch errors before a paper is shared publicly, and to conduct large-scale audits of the published literature (e.g., Nuijten & Polanin, 2020). A field that prioritizes self-correction is one that invests in developing and using these kinds of tools, and that is shown to have a low rate of errors when these tools are applied to its published literature.

Software such as reference managers can also check whether any of the references in a manuscript or published article are retracted papers, using the Crossref API and/or the Retraction Watch database (Abritis, Marcus, & Oransky, 2021). Citing a retracted paper is not necessarily an error, but it is poor practice to cite such a paper without noting that it was retracted, so a high rate of uncritical citations to retracted papers is a sign that correction efforts in a field are not having the appropriate effect.

Image manipulation is another 'error' that could potentially be detected using automated algorithms. In biology, private "paper mills" are paid by researchers to fabricate data and write fraudulent papers to submit to journals. Many of these papers contain manipulated images that traditionally are flagged by human experts (Else & van Noorden, 2021), but a group of publishers has come together with the experts to develop automated detection efforts (van Noorden, 2020). Investment in such tools is a sign of a scientific community's commitment to self-correction.

Computational reproducibility checks

A particularly thorough sort of error detection is computational reproducibility checks, which test whether the statistical results reported in a paper can be computationally reproduced using the original data. The vast majority of scientific results involve numerical data analysis to calculate summary and test statistics. Both the analyses themselves and the reporting of the associated numbers are quite error-prone. There is a large gap between the raw data and the claims made about it in a paper; as Buckheit and Donoho (1995) wrote, "An article [...] in a scientific publication is not the scholarship itself, it is merely advertising for the scholarship." (p. 5). To ensure that errors and discrepancies between the data and the report are caught, it is crucial that audits be performed to compare the reports to the underlying data.

The data and code underlying the quantitative results of scientific outputs have become increasingly available in many fields. This has made it possible for meta-scientists to systematically audit samples of published papers to evaluate whether the papers are computationally reproducible. Such efforts have yielded disappointing results. In one analysis

of 37 papers published in the journal *Cognition* after the journal implemented a mandatory data sharing policy, the meta-researchers were only able to computationally reproduce all of the results in 11 out of 35 articles with reusable data (Hardwicke et al., 2018). A further 11 could be reproduced after getting assistance from the original authors. For the remaining 13 articles, the full results could not be reproduced from the original data, even after obtaining assistance from the original authors.

There is a great deal of variance in how seriously journals and researchers take the problem of computational reproducibility, and therefore how effortful it is to check the computational reproducibility of a set of papers. Many journals continue to leave it up to authors to decide whether or how to share their data and code with other researchers. This can make checking computational reproducibility extremely time-consuming. Other journals, such as the *American Journal of Political Science*, require authors to submit reproducibility "packages" which are then checked by an external team, and any problems must be fixed before the article can be published (https://ajps.org/ajps-verification-policy/). The journal *eLife* now uses technology that allows published articles to have interactive figures with live code that connects the data to the analyses and resulting statistics, all within the browser (Maciocci et al., 2019). Some fields, such as parts of high-energy physics, have a tradition of setting up their full analysis pipelines prior to feeding it any real data, and running tests on the pipeline to check for errors. Software tools such as Docker (Boettiger, 2015; Clyburne-Sherin et al, 2019) allow anyone to archive a fully working version of their analysis code. We can therefore evaluate a scientific community's commitment to self-correction by evaluating its journals' policies, or conducting post-publication reproducibility checks and directly estimating the computational reproducibility of published findings. Both the difficulty of conducting such audits, as well as the success rate of computational reproducibility audits, tell us something about how seriously a field takes error detection and error control.

Empirical replicability checks

Another important role of critical appraisal is to identify and correct erroneaous statistical conclusions. Erroneous conclusions can be drawn about the size of an effect, the sign of the effect, and/or whether the effect exists at all (i.e., is meaningfully different from zero). These errors can come about by random chance, accidents (e.g., miscoding experimental conditions; Strand, 2020), or questionable research practices such as p-hacking (Simmons, Nelson, & Simonsohn, 2011). Anonymous surveys of researchers in criminology (Chin, Pickett, Vazire, & Holcombe, in press), psychology (John, Loewenstein, & Prelec, 2012), political science (Franco, Malhotra, & Simonovits, 2015), management (O'Boyle et al., 2017), education (Makel et al., 2021), quantitative communication (Bakker et al., 2020), and ecology and evolutionary biology (Fraser et al., 2018) all indicate that substantial proportions of researchers admit to engaging in questionable research practices.

The only sure way to estimate the prevalence of statistical inference errors (false positives, false negatives, inflated effect sizes, etc.) is a very expensive and time-consuming process: replicating studies by using similar methods to collect new data, while reducing the opportunities for bias. Camerer et al. (2018) replicated 21 systematically-selected experimental social science studies published in *Nature* and *Science* between 2010 and 2015, with much larger sample sizes than the original design, but disappointing results – only 62% yielded the same statistically significant

result as the original, and effect sizes were often substantially smaller. Such efforts should themselves be repeated to help assess whether fields are improving in whether their findings replicate.

Replication studies of individual papers are increasingly welcomed by journals and the community (at least in policy, if not always in practice), and this trend must continue if we want people to have confidence that most results in a field or literature are replicable. Thus, the prevalence of replication studies, their appearance in high profile outlets, their impact, and a field's receptiveness to high-quality replication projects and their results are all indicators of a field's commitment to self-correction.

Pre-publication peer review

Across all scientific fields, peer-reviewed scientific journals conduct some form of peer review before accepting papers. This is not always done prior to publication (e.g., overlay journals conduct the peer review after a preprint has been posted), but what such journals have in common is that they give feedback to the paper's authors at a stage that is early enough that authors are likely to consider changing their paper in light of the feedback. For lack of a better phrase, we will call this type of peer review "pre-publication peer review". Scientific fields differ in how careful and effective their journals' pre-publication peer review process is.

A journal that is committed to enabling a self-correcting science should take steps to ensure that its peer review system is effective, fair, and accountable. To this end, journals can develop and test policies that encourage reviewers and editors to identify critical errors and flaws, such as requiring authors to share materials, data, and code necessary to reproduce and replicate the work, or developing and validating rubrics that ask reviewers and editors to identify potential errors and flaws such as lack of transparency, irreproducibility, or inappropriate design, analyses, or conclusions.

The Registered Reports article format that has been adopted by many journals (Chambers et al., 2015) encourages scientists to identify and correct errors early in the research process. A Registered Reports must be submitted to the journal before data collection. That way, reviewers and editors can provide feedback on the design and execution of the study when changes can still be implemented. Registered Reports reduce the opportunity for authors and editors to engage in the selective reporting and selective publication of results – practices that have led some scientific literatures to contain a high rate of false positives.

A second thing journals can do to increase the effectiveness of peer review is create incentives for scientists to invest in critically reviewing each other's work, for example by paying reviewers or recognizing them for providing high-quality reviews.

This brings us to the third thing journals can do: make the process and outputs of peer review as transparent as possible (i.e., publish the content of reviews and decision letters). In addition to giving reviewers the opportunity to build a reputation for high quality critiques, this also allows readers to evaluate the journal's process, as well as benefit from the knowledge produced during pre-publication peer review. Some journals, such as *Meta-Psychology*, publish all of their peer reviews.

Finally, journals can conduct internal audits of how well their peer review policies are being put into practice. Such checks of whether policies are being followed, and whether reviews are high quality can also reveal signs of bias or corruption (e.g., O'Grady, 2020).

In many fields, most journals have done little to ensure that they effectively detect and correct errors in submitted manuscripts. The TOP Factor (Mayo-Wilson et al., 2021) is a journal-level metric that rates journals' policies on the degree to which they promote transparent and open reporting of research. As of this writing, only 8 of 758 journals score 20 or above out of a possible 30 points. Some specific policies, however, such as accepting Registered Reports, have been taken up quite rapidly – 293 journals publish Registered Reports as either a regular article type or as part of a single special issue. Other practices, such as paying reviewers or publishing the peer review histories of papers, seem to be quite rare, although we don't know of systematic data on this.

At its worst, the journal-based peer review process can introduce new errors or biases, for example by encouraging authors to drop studies or analyses that dilute the potential impact of the paper. They can also perpetuate eminence bias, by treating papers from famous authors more favorably. If journals do not become a more positive force for self-correction in science, the scientific community can take matters into their own hands and cultivate platforms to deliver early feedback to authors to catch and fix flaws in scientific papers.

The growth of preprints has been a boon to pre-(journal)-publication discussion of research. Many researchers discuss preprints on social media (e.g., Twitter), and some authors use those discussions to inform their revisions of their manuscripts. However, platforms such as Twitter are not optimized for sober and even-handed discussions.

Fortunately, platforms designed specifically for scientific discussions have been growing in popularity. Prereview.org was created for the biological sciences and appears to have deployed effectively for COVID-19 research. Pubpeer.com has been used to highlight image duplication and other issues with published research, contributing to many retractions in the past, and now hosts a growing number of discussions of preprints. PreLights (https://prelights.biologists.com/about-us/) is a community of early-career biologists who comment on biology preprints at a dedicated website and include responses from authors. In short, whether through traditional journals or via more grassroots efforts, a field that is committed to scientific self-correction will find ways to detect and correct errors in early drafts of manuscripts.

Post-publication peer review

The growing popularity of preprints is blurring the line between pre-publication and post-publication peer review. We discuss what a healthy system of early feedback to authors looks like in the section above ("Pre-publication peer review"). Here, we use the term "post-publication peer review" to refer to critiques that are not aimed primarily at helping authors to improve the critiqued paper, but whose main goal is to point out errors or flaws in published work to the broader readership and to inform further work.

A field should not have a very high bar for publication of critical comments, and such comments should be easy to find. This is important because, as Heathers (2015) wrote, "Formal critical

correspondence is the ONLY way to make a visible and public correction to the official version of a journal article. It is written, archived and maintained in such a manner that it is irreparably bound to the criticism, which also received the distinction of being 'published'." Anecdotally, however, there is a high bar in many fields (see the "Traditional scientific institutions are ineffective self-correction mechanisms" section above).

Fortunately, reports of errors can now be shared on the internet without any gatekeeping, such as using web annotations (e.g. with hypothes.is). However, the internet is a big and rather untrustworthy place, so for these reports to come to the attention of researchers on the topic, a curated database must link the comments to the paper (Eagleman & Holcombe, 2003). PubPeer is one website that provides this service, and thriving communities in some fields, such as cancer biology, have begun to embrace it as a place to discuss possible errors and data issues in specific papers. Additional venues have flowered in the wake of the COVID19 pandemic and the associated need to rapidly review new research (Holcombe, 2020). These platforms serve an important function, although they are unlikely to soon achieve the level of effectiveness of the publication of critical commentaries in journals. Thus, the highest level of commitment to self-correction in the domain of post-publication peer review would be an abundance of critical commentaries published in high-profile journals. A less compelling but still important indicator would be the amount and quality of critical commentary happening in less traditional channels.

The existence of published criticism does not seem to be sufficient for a field to self-correct. In a previous section, we mentioned the recent study by Serra-Garcia & Gneezy (2021) of top psychology, economics, and general science journals that publish social science. The study focused on findings that subsequently failed to replicate, found that such articles are cited more than others, and that the publication of a failure to replicate did not change the citation advantage. After the failed replication, only 12% of subsequent articles that cited the original article acknowledged the replication failure. Evidently, the social sciences do not respond appropriately to failed replications. This is consistent with our suspicion that at least some of them do not practice a healthy level of self-correction.

Detecting errors and biases in a literature

In addition to critiquing individual papers, a field committed to self-correction will look for concerning patterns across the field. Systematic reviews and meta-analyses are often considered to be the highest level of evidence, but often these tools gloss over flaws in individual studies, rather than to look for signs of systematic problems, and this can further amplify any biases (Kvarven, Strømland, & Johannesson, 2020).

Techniques to detect publication biases and other sources of distortion in a field are evolving rapidly (Simonsohn et al., 2014; Andrews & Kasy, 2019). Some of these techniques use the statistical tests associated with key results across a set of papers to look for patterns that are unlikely to happen in the absence of bias. For example, the distribution of *p*-values among statistically significant results should be heavily right-skewed (many *p*-values close to zero, very few that are barely significant) when effects are real and there is no systematic bias. The distribution of a collection of statistically significant *p*-values associated with a set of key findings can be checked for the appropriate level of right skew. A relative lack of right skew suggests that the set of statistical results reflects bias, and we should be careful interpreting these studies or meta-analyses based on this literature. Such techniques are now regularly applied to

groups of papers in systematic reviews and meta-analyses, and can also be used for other metascientific purposes such as examining levels of bias across journals, over time, or across different methods or practices (e.g., papers reporting primary analyses with vs. without covariates). These tools help us to monitor how a field is doing at combating publication bias and other distorting statistical practices, such as *p*-hacking. Moreover, by examining where bias is most severe, each field can begin to develop interventions for reducing statistical bias and improving the credibility of its published literature.

Another test of bias that can be applied at the level of a literature is the dearth of negative or null results. Multiple fields of science have a very poor record of publishing negative and null results (Sterling, Rosenbaum, and Weinkam, 1995; Fanelli, 2010; 2012; Greenwald, 1975; Hubbard & Armstrong, 1992; Franco, Malhotra, & Simonovits, 2014; Sterling, 1959; Sterling, Rosenbaum, & Wenkam, 1995), suggesting that the published literature is biased and incomplete. Fortunately, many researchers are already tracking this indicator of the credibility of specific fields. Metascientists are developing new techniques for estimating publication bias (e.g., Andrews & Kasy, 2019) and initiatives such as Registered Reports, described above, are being adopted to combat it (Chambers et al., 2015). Preliminary results from analyses of the emerging Registered Reports literature suggest that Registered Reports are successful at allowing more negative results to make it into the published literature (Allen & Mehler, 2019; Scheel, Schijen, & Lakens, 2020).

While most efforts aimed at evaluating bias in a literature are currently focused on statistical biases, a few fields have developed tools to detect evidence for other systematic errors and biases. In the biomedical literature, Critical Appraisal Tools (CATs) are often used to evaluate the quality of individual studies being considered in systematic reviews. In the social sciences, Levy Paluck and colleagues (Levy Paluck, Porat, Clark, & Green, 2020) recently reviewed the literature on prejudice reduction and identified several concerning patterns in the types of designs and methods used, and conclusions drawn.

In a field that is committed to self-correction, we would expect to see widespread development and application of tools to evaluate the studies going into systematic reviews and meta-analyses. Moreover, we would expect these tools to show low levels of error and bias, and a great deal of concern and effort to redress the situation when bodies of literature are found to be problematic.

Strong theory and predictions

Parts of physics and some other fields are guided by theories that make precise predictions. Einstein's theory of general relativity predicted that light would bend twice as much as Newtonian theory predicted. When Einstein's prediction was confirmed by observation (Dyson, Eddington, & Davidson, 1920), his theory became much more popular; it became highly credible.

Construction of such highly predictive and quantitative theories may not be appropriate in some sciences, such as some areas of psychology, which arguably should concentrate on characterizing phenomena (Perfors, 2020; Rozin, 2009). However, in areas where construction of predictive theories is possible, fields can reach greater consensus about what empirical results mean for the success of different theories. The best of those theories can accumulate a track record of accurate predictions, which should increase the credibility of the associated fields. The adoption of

preregistration can facilitate accumulating such a track record. One reason, we believe, that the role of theory in social sciences is hotly contested (Meehl, 1990; Yarkoni & Westfall, 2017; Fried, 2020; Cummins, 2000; van Rooij & Baggio, 2020) is because we have so little unbiased data, and little agreement, on where and when theories make good predictions. It may be that some fields already contain quantitative theories that have been making successful predictions, but the absence of preregistration as a practice has made this hard to recognize.

Diversity

Diversity within a field helps overcome hindrances such as shared biases and close ties among the researchers studying a particular topic (Longino, 1990). People with different theoretical commitments and knowledge bases are likely to catch different sorts of errors. Those attached to different theories will be motivated to check, or double-check, different calculations. In other words, diversity within a scientific community helps us identify biases or blind spots in each other's research, thus making us, collectively, more objective. In contrast, uniformity in approaches and perspectives breeds group-think. As Cordelia Fine (2020) has writtten, "Whether for reasons self-serving or benign, everyone comes laden with prior knowledge, background assumptions and frameworks. That's why it takes a diverse village, so to speak, to nurture scientific objectivity."

Diversity in personal and institutional ties may also be important. In some areas of psychology research, nearly all researchers who conduct studies with a particular paradigm may know each other. The gatekeepers at certain journals may be exclusively drawn from narrow clans, which can result in the suppression of new perspectives and approaches. This would be difficult to quantify, but qualitative studies and network analyses might provide some indication of whether this is the case for a particular research area. Geographical and institutional diversity in journal editors can be an additional, albeit very imperfect, proxy. More broadly, diversity in personal experiences (e.g., because of one's ethnic or racial identity, gender, or income level) is likely to lead to a broader range of perspectives on what research questions are important, what methods are appropriate, and how evidence should be interpreted. A demographically homogenous field is likely to have significant blind spots in what they study and how they study it.

Diversity is unique in its importance to scientific self-correction because it enhances the effectiveness of all other critical appraisal mechanisms, such as those described above. For example, the type of errors that a reviewer is likely to look for and detect in pre- or post-publication peer review likely depends in part on their background, training, intellectual commitments, experiences, and values. To the extent that a scientific community includes people who are diverse in these characteristics, their peer review processes will be more robust. More generally, a more diverse scientific community is likely to develop techniques for detecting and correcting a more diverse, and therefore more comprehensive, range of scientific errors.

Conclusion

A high level of trust is justified when something has a long and consistent track record of claims being true. The replication crisis shows that this is unfortunately not the case for some areas of science. The American president Ronald Reagan, in negotiation with his Soviet counterpart Mikhail Gorbachev, frequently invoked a Russian proverb that means "trust, but verify" ("Doveryai, no proveryai"). At one time, perhaps, there was little reason to suspect that the

sciences could not be trusted and that verification was necessary. The replication crisis, along with the relatively new phenomenon of scientists being spectacularly wrong in public (as we saw with some high-profile errors in COVID-19 research), have highlighted the crucial role of verification in buttressing the credibility of science.

To justify trust in today's science, we should institute practices that facilitate assessment of whether trust is warranted. We can no longer blithely declare "science is self-correcting" and expect that to reassure science's stakeholders. We need to determine where self-correction mechanisms are working well and working rapidly so that we can strengthen public trust in science where it is warranted, improve the self-correcting mechanisms where they are weak, and prevent less robust corners of science from dragging other areas down with them.

References

Allen, C., & Mehler, D. M. (2019). Open science challenges, benefits and tips in early career and beyond. *PLoS biology*, *17*(5), e3000246.

Anderson, M. S., Martinson, B. C., & De Vries, R. (2007). Normative dissonance in science: Results from a national survey of US scientists. *Journal of Empirical Research on Human Research Ethics*, *2*(4), 3-14.

Andrews, I., & Kasy, M. (2019). Identification of and Correction for Publication Bias. American *Economic Review, 109(8),* 2766–2794. https://doi.org/10.1257/aer.20180310

Abritis, A., Marcus, A., & Oransky, I. (2021). An "alarming" and "exceptionally high" rate of COVID-19 retractions? *Accountability in Research*, *28*(1), 58–59. https://doi.org/10.1080/08989621.2020.1793675

Andrews, I., & Kasy, M. (2019). Identification of and Correction for Publication Bias. *American Economic Review*, *109*(8), 2766–2794. https://doi.org/10.1257/aer.20180310

Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Aken, M. A. G. van, Weber, H., & Wicherts, J. M. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality, 27(2),* 108–119. https://doi.org/10.1002/per.1919

Bakker, B. N., Jaidka, K., Dörr, T., Fasching, N., & Lelkes, Y. (2020, November 18). Questionable and open research practices: attitudes and perceptions among quantitative communication researchers. https://doi.org/10.31234/osf.io/7uyn5

Boettiger, C. (2015). An introduction to Docker for reproducible research. ACM SIGOPS Operating Systems Review, 49(1), 71-79.

Brown, N. J., & Heathers, J. A. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science, 8(4),* 363–369.

Buckheit, J. B., & Donoho, D. L. (1995). Wavelab and reproducible research. In *Wavelets and statistics* (pp. 55-81). Springer, New York, NY.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2(9),* 637–644. https://doi.org/10.1038/s41562-018-0399-z

Campbell, D. T. (1988). *Methodology and epistemology for social sciences: Selected papers*. University of Chicago Press.

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: Realigning incentives in scientific publishing. *Cortex, 66,* A1–A2.

Chin, J.M., Pickett, J.T., Vazire, S., & Holcombe, A.O. (in press). Questionable Research Practices and Open Science in Quantitative Criminology. *Journal of Quantitative Criminology*.

Christensen, G., Wang, Z., Levy Paluck, E., Swanson, N., Birke, D., Miguel, E., & Littman, R. (2020). Open science practices are on the rise: The State of Social Science (3S) Survey. MetaArXiv. https://osf.io/preprints/metaarxiv/5rksu/

Clyburne-Sherin, A., Fei, X., & Green, S. A. (2019). Computational Reproducibility via Containers in Psychology. *Meta-Psychology, 3*.

Cummins, R. (2000). "How does it work?" versus "What are the laws?": Two conceptions of psychological explanation. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and Cognition* (p. 117–144). Cambridge, MA: The MIT Press.

Dyson, F. W.; Eddington, A. S.; Davidson C. (1920). "A determination of the deflection of light by the Sun's gravitational field, from observations made at the total eclipse of 29 May 1919". *Philosophical Transactions of the Royal Society, 220A* (571–581): 291–333.

Eagleman, D. M., & Holcombe, A. O. (2003). Improving science through online commentary. *Nature, 423(6935),* 15–15.

Else, H., & Noorden, R. V. (2021). The fight against fake-paper factories that churn out sham science. *Nature*, *591*(7851), 516–519. https://doi.org/10.1038/d41586-021-00733-5

Errington, T. (2019, September). Talk presented at the Metascience Symposium, Palo Alto, CA. Retrieved from https://www.metascience2019.org/presentations/tim-errington/

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). Science forum: An open investigation of the reproducibility of cancer biology research. *Eife*, *3*, e04333.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PloS one*, *5*(4), e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891-904.

Fine, C. (2020, July). Sexual dinosaurs: The charge of 'feminist bias' is used to besmirch anyone who questions sexist assumptions at work in neuroscience. *Aeon.* Retrieved from https://aeon.co/essays/trumped-up-charges-of-feminist-bias-are-bad-for-science

Fiske, S. T., Schacter, D. L., & Taylor, S. E. (2016). Introduction. *Annual Review of Psychology, 67*.

Franco, A., Malhotra, N., & Simonovits, G. (2015). Underreporting in political science survey

experiments: Comparing questionnaires to published results. *Political Analysis, 23,* 306-312.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345(6203),* 1502–1505. https://doi.org/10.1126/science.1255484

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fiddler, F. (2018). Questionable research practices in ecology and evolution. *PLoS One, 13*(7): e0200303.

Friedman, H., MacDonald, D. A., & Coyne, J. (in press). Working with psychology journal editors to correct problems in the scientific literature. *Canadian Psychologist.*

Funk, C., Hefferon, M., Kennedy, B., & Johnson, C. (2019). Trust and mistrust in American's views of scientific experts. Pew Research Center. https://www.pewresearch.org/science/2019/08/02/trust-and-mistrust-in-americans-views-of-scientific-experts/

Gelman, A. (2017). Ethics and statistics: Honesty and transparency are not enough. *Chance*, *30*(1), 37-39.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641-651.

Goldacre, B., Drysdale, H., Dale, A., Milosevic, I., Slade, E., Hartley, P., Marston, C., Powell-Smith, A., Heneghan, C., & Mahtani, K. R. (2019). COMPare: A prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, *20*(1), 118. https://doi.org/10.1186/s13063-019-3173-2

Greenwald, A. G. (1975), "Consequences of prejudice against the null hypothesis," Psychological Bulletin, 82, 1- 20.

Guston, D. H. (1994). The demise of the social contract for science: Misconduct in science and the non-modern world. *The Centennial Review, 38(2),* 215–248. https://www.jstor.org/stable/23740126

Haack, S. (2003). *Defending Science - Within Reason: Between Scientism and Cynicism.* Prometheus Books.

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., ... & Lenne, R. L. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society open science*, *5*(8), 180448.

Heathers, J. [@jamesheathers]. (2019, March 01). In reality, mechanisms to correct bad science are slow, unreliably enforced, capricious, run with only the barest nod towards formal policy, confer no reward and sometimes punitive elements for a complainant who might use them. [Tweet]. Retrieved from https://twitter.com/jamesheathers/status/1101161838308401157

Heathers, J. (2015, October 2). A General Introduction: Formal Criticism in Psychology. Medium. https://medium.com/@jamesheathers/a-general-introduction-formal-criticism-in-psychology-ba193a940ec8

Heesen, R., & Bright, L. K. (n.d.). Is Peer Review a Good Idea? *The British Journal for the Philosophy of Science.* https://doi.org/10.1093/bjps/axz029

Holcombe, A. O. (2020, May). As new venues for peer review flower, will journals catch up? Psychonomic Society featured content. Retrieved from: https://featuredcontent.psychonomic.org/as-new-venues-for-peer-review-flower-will-journals-catch-up/

Hubbard, R., & Armstrong, J. S. (1992), "Are null results becoming an endangered species in marketing?" *Marketing Letters, 3,* 127-136.

Jamieson, K. H., McNutt, M., Kiermer, V., & Sever, R. (2019). Signaling the trustworthiness of science. *Proceedings of the National Academy of Sciences*, 116(39), 19231–19236. https://doi.org/10.1073/pnas.1913039116

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable

research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.

Maciocci, G., Aufreiter, M. and Bentley, N. (2019). Introducing eLife's first computationally reproducible article. *ELife* blog. https://elifesciences.org/labs/ad58f08d/introducing-elife-s-first-computationally-reproducible-article

Makel, M. C., Hodges, J., Cook, B. G., & Plucker, J. (2021). Both Questionable and Open Research Practices Are Prevalent in Education Research. *Education Researcher*, 1-12.

Markman, A. Why science is self-correcting. (2010). Ulterior motives blog on Psychology Today. Retrieved July 28, 2020, from http://www.psychologytoday.com/blog/ulterior-motives/201008/why-science-is-self-correcting

Mayo-Wilson, E., Grant, S., Supplee, L., Kianersi, S., Amin, A., DeHaven, A., & Mellor, D. T. (2021). Evaluating implementation of the Transparency and Openness Promotion Guidelines: The TRUST Process for rating journal policies, procedures, and practices. MetaArXiv. https://doi.org/10.31222/osf.io/b3wju

McConnell, J. [@JohnSMcConnell] (2020, June 14). Retraction should be reserved for publication misconduct. Otherwise, let the self-correcting mechanisms of science take their course. [Tweet]. Retrieved from https://twitter.com/JohnSMcConnell/status/1271860082427547649

Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. doi:10.2466/pr0.1990.66.1.195

National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115(11),* 2600–2606.

Nuijten MB, Hartgerink CHJ, van Assen MA, Epskamp S, Wicherts JM (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods*, 48: 1205–1226.

Nuijten, M. B., & Polanin, J. R. (2020). "statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Research Synthesis Methods*, *11*(5), 574–579.

O'Boyle, E. H, Jr., Banks, G. C., & Gonzalez-Mulé, E. (2017). The chrysalis effect: How ugly initial results metamorphize into beautiful articles. *Journal of Management, 43*(2), 376-399.

O'Grady, C. (2020, October). Delete offensive language? Change recommendations? Some editors say it's ok to alter peer reviews. *Science.* doi: 10.1126/science.abf4690

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

Oreskes, N. (2019). *Why trust science?*. Princeton University Press.

Perfors, A. (2020). Generating theories is hard, and none of our theories are good enough. Presentation for the Annual Meeting of the Society for Mathematical Psychology. https://virtual.mathpsych.org/presentation/109

Pickett, J. T. (2020). The Stewart Retractions: A Quantitative and Qualitative Analysis. *Econ Journal Watch, 17(1),* 152–190.

Retraction Watch (2014, March). In sharp resignation letter, former ORI director Wright criticizes bureaucracy, dysfunction. Retrieved from https://retractionwatch.com/2014/03/13/in-sharp-resignation-letter-former-ori-director-wright-criticizes-bureaucracy-dysfunction/

Rozin, P. (2009). What Kind of Empirical Research Should We Publish, Fund, and Reward? A Different Perspective. *Psychological Science, 4(4),* 435–439.

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, *7*(21), eabd1705. https://doi.org/10.1126/sciadv.abd1705

Scheel, A. M., Schijen, M., & Lakens, D. (2020). An excess of positive results: Comparing the standard Psychology literature with Registered Reports. Retrieved from https://psyarxiv.com/p6e9c/

Schroter, S., Black, N., Evans, S., Godlee, F., Osorio, L., & Smith, R. (2008). What errors do peer reviewers detect, and does training improve their ability to detect them? *Journal of the Royal Society of Medicine, 101(10),* 507–514. https://doi.org/10.1258/jrsm.2008.080062

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143(2),* 534.

Smith, T. (2004). "Social" objectivity and the objectivity of values. In *Science, Values, and Objectivity.* Peter Machamer Gereon Wolters, eds. University of Pittsburgh Press.

Sterling, T. D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.2307/2282137

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995), "Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa," *The American Statistician, 49,* 108- 112.

Strand, J. (2020, March). Scientists make mistakes. I made a big one. Retrieved from https://elemental.medium.com/when-science-needs-self-correcting-a130eacb4235

Sztompka, P. (2007). Trust in science: Robert K. Merton's inspirations. *Journal of Classical Sociology*, *7*(2), 211-220.

van Rooij, I. & Baggio, G. (2020) Theory Development Requires an Epistemological Sea Change. *Psychological Inquiry*, 31:4, 321-325, DOI: 10.1080/1047840X.2020.1853477

Tenney, E. R., Costa, E., Allard, A., & Vazire, S. (2020). Open science and reform practices in organizational behavior research over time (2011 to 2019). Manuscript in preparation.

Van Noorden, R. (2020). Publishers launch joint effort to tackle altered images in research papers. *Nature News*. https://doi.org/10.1038/d41586-020-01410-9

Vazire, S. (2020, January). Do we want to be credible or incredible? *APS Observer*. Retrieved from https://www.psychologicalscience.org/observer/do-we-want-to-be-credible-or-incredible

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Ziemann, M., Eren, Y., & El-Osta, A. (2016). Gene name errors are widespread in the scientific literature. *Genome biology*, *17*(1), 1-3.